Investigating students' behavior and performance in online conceptual assessment

Bethany R. Wilcox and Steven J. Pollock
Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA

(Received 16 August 2019; published 2 December 2019)

Historically, the implementation of research-based assessments (RBAs) has been a driver of educational change within physics and helped motivate adoption of interactive engagement pedagogies. Until recently, RBAs were given to students exclusively on paper and in class; however, this approach has important drawbacks including decentralized data collection and the need to sacrifice class time. Recently, some RBAs have been moved to online platforms to address these limitations. Yet, online RBAs present new concerns such as student participation rates, test security, and students' use of outside resources. Here, we report on a study addressing these concerns in both upper-division and lower-division undergraduate physics courses. We gave RBAs to courses at five institutions; the RBAs were hosted online and featured embedded JavaScript code which collected information on students' behaviors (e.g., copying text, printing). With these data, we examine the prevalence of these behaviors, and their correlation with students' scores, to determine if online and paper-based RBAs are comparable. We find that browser loss of focus is the most common online behavior while copying and printing events were rarer. We found that correlations between these behaviors and student performance varied significantly between introductory and upper-division student populations, particularly with respect to the impact of students copying text in order to utilize internet resources. However, while the majority of students engaged in one or more of the targeted online behaviors, we found that, for our sample, none of these behaviors resulted in a significant change in the population's average performance that would threaten our ability to interpret this performance or compare it to paper-based implementations of the RBA.

DOI: 10.1103/PhysRevPhysEducRes.15.020145

I. INTRODUCTION AND MOTIVATION

Research-based assessments (RBAs) have become a cornerstone of physics education research (PER) due in large part to their ability to provide a standardized measure of students' learning that can be compared across different learning environments or curricula [1]. As such, these assessments are a critical step along the path towards making evidenced-based decisions with respect to teaching and student learning. RBAs have historically been a strong driver in promoting the need for, and adoption of, educational reforms in undergraduate physics courses (e.g., Refs. [2–4]). It can be argued that, without the invention and consistent use of RBAs, the PER community might not have the same focus on active learning and interactive engagement that is does today.

However, despite their value, there are a number of barriers to wide-scale implementation of RBAs that stand in the way of their integration into physics departments [5,6]. For example, most of the existing RBAs require that an instructor sacrifices 1–2 full class periods to administering the RBA pre-instruction and postinstruction. For many instructors feeling pressure to cover as much content as possible over the course of a semester, this sacrifice is difficult to justify. In addition to the demand for class time, instructors must also sacrifice valuable time outside of class to analyze their students' performance. Many instructors are not experts in assessment and struggle with analysis and interpretation of their students' scores. This can make faculty particularly reluctant to sacrifice class time to an assessment that they are ultimately unable to identify actionable results from.

Recently, physics education researchers have attempted to address both of these challenges by shifting RBAs to online platforms (e.g., Refs. [5,7–9]). Hosting the RBAs online allows instructors to assign the RBA for students to complete outside of class, freeing them from the need to sacrifice class time. Moreover, the online platform allows for easy standardization and centralization of the data collection and analysis process. This has two major advantages for the instructor. By automating the analysis of students' responses, these centralized systems make it so that the instructor no longer needs to perform this analysis themselves. Moreover, centralizing data collection ensures

the aggregation of comparison data that can be used to facilitate meaningful comparisons and can help instructors, and researchers, to identify actionable implications from their students' performance. However, while these online systems have a lot of potential for encouraging more widespread use of RBAs by removing barriers to their use, these systems bring with them an number of other concerns, particularly around the potential for reduced participation rates, students' use of outside resources, potential for distraction, and breaches of test security [6].

Here, we build on prior work (see Sec. II) to investigate the extent to which these concerns factor into students' scores when completing standardized physics conceptual assessments online. We include data from both introductory and upper-level contexts as there are significant differences between the student populations at these two levels [10], which could have implications for how students engage with an online assessment. In the next section (Sec. II), we describe prior work around online conceptual assessments. We then discuss the context and methods used in this study (Sec. III), and present our findings with respect to students' online behaviors when taking the RBAs as well as how these behaviors correlate with their overall performance (Sec. IV). Finally, we end with a discussion of our conclusions, limitations, and implications of the study (Sec. V).

II. BACKGROUND

Significant prior work has been done to address some of the concerns around online conceptual assessment as part of the Learning Assistant Student Supported Outcomes (LASSO) study. Specifically, they have investigated concerns about changes in scores and participation rates between online and paper-based administrations of the Force Concept Inventory (FCI) [11] and Conceptual Survey of Electricity and Magnetism (CSEM) [12] in the context of introductory courses. They found that, when looking at all courses in aggregate, participation rates tended to be lower for online RBAs [13,14]. However, this difference between the two formats vanished when best practices were used for the online implementations. The best practices identified in the LASSO study include multiple email and in-class reminders, and offering participation credit for completing both the pretests and post-tests. Moreover, they also found that, when participation rates were similar, students' overall performance was also statistically comparable [15].

Historically, multiple researchers have addressed the issue of the comparability of online assessments both within and outside of PER. For example, MacIsaac *et al.* [6] also found no difference in students' scores on the FCI between web-based and paper-based administrations. In addition to investigating overall score, they also saw no difference in performance on individual items and no difference based on students' gender. However, while studies within PER have consistently indicated that there is no difference in performance between online and paper-based RBAs, the results from outside PER are more varied. Many studies have documented no statistically significant difference between students' performance on online multiple-choice tests (e.g., Refs. [16,17]), while others have reported cases where online tests scored statistically higher or lower than associated paper-based tests (see Refs. [17,18] for reviews). The variation in these studies has led to the recommendation that, while online and paper-based tests *can* be equivalent, it should not be *assumed* that they are equivalent until it has been clearly demonstrated that they actually are [18].

A smaller body of work has focused specifically on investigating the validity of concerns about students' use of outside resources (e.g., the internet or other students) or breaches in test security. For example, Haney and Clarke [19] collected timing and path data from students who took a series of online quizzes over the course of a one semester course. By analyzing patterns in students' responses (e.g., similarities in two students' response patterns combined with close timing of the two submissions) to identify likely cases of students collaborating on the assignments. They found that this type of collaboration increased as the semester went on and students adapted to the online quiz format. They also asked students to self-report whether they collaborated with others for the quizzes and found students reported collaboration with similar frequencies to what was detected in the response patterns.

Another study conducted in the context of an introductory astronomy course looked at different online student behaviors when taking an online conceptual assessment [20]. In this study, Bonham [20] used JavaScripts and other applets to detect when students engaged in behaviors like printing browser pages, coping or highlighted text, and switching in to other browser windows while taking an online astronomy concept assessment. They found no instances of students printing pages, and only 6 cases (out of 559) they deemed were probable incidence of students copying text. Students switching browser windows was more common; however, Bonham argued these events appeared random and were not systematically associated with particular questions. There were several important limitations to Bonham's study. In browsers other than Microsoft Explorer, copy events and save events were detected through the proxies of highlighting text and page reloads, respectively. As Bonham noted, highlighting text as a proxy for copying results in many false positives, and there was no discussion of how these behaviors related to performance on the RBA. Here, we replicate and extend the study by Bonham in the context of physics courses at both the upper-division and introductory levels.

III. CONTEXT AND METHODS

Four different physics RBAs were used in this study—two upper division and two introductory. The two upper-division RBAs used in this study were the Quantum

Mechanics Conceptual Assessment (QMCA) [21] and the Colorado Upper-division Electrostatics Diagnostic (CUE) [22]. Both the QMCA and CUE are multiple-choice or multiple-response assessments targeting content from the first semester in a two-semester sequence in junior-level quantum mechanics, and electricity and magnetism, respectively. The two introductory assessments used were the Force and Motion Conceptual Evaluation (FMCE) [23] and the Brief Electricity and Magnetism Assessment (BEMA) [24]. The FMCE and BEMA are both multiple-choice assessments targeting content from the first and second semester, respectively, of a two-semester, calculus-based introductory physics course. All four RBAs were administered online, using the survey platform Qualtrics, during the final week of the regular semester. The online versions of the RBAs were designed to mirror the paper versions as faithfully as possible. For example, each separate page on the paper versions was offered as a separate browser page in the online version. Students could also navigate freely both forward and backward within the assessment, as they would be able to with a paper exam packet.

Student responses were collected from 2 introductory courses and 10 upper-division courses at eight institutions. All eight institutions are four-year universities spanning a range of types including three doctoral-granting institutions classified as very high research, two master's-granting institutions (one classified as Hispanic-serving), and three bachelor's-granting institutions. The authors taught two of the upper-division courses, and the remaining instructors volunteered. Data from the two introductory courses came only from the doctoral-granting institutions. In all cases, the instructors offered regular course credit to their students for simply completing the RBA (independent of performance). In most cases, students received multiple in-class reminders to complete the assessment. After elimination of responses that were marked as invalid (e.g., due to too many blanks; less than 3% of nonduplicate responses were identified as invalid), participation rates by course varied from 70% to 100% over all 12 courses. Criteria for identifying a response as invalid were failure to submit the assessment within 1 week of starting it or leaving 6 or more questions blank. In the interests of establishing a "worst case scenario" measure of the online implementations, we did not impose a minimum time threshold for valid submissions. Since the goal of this study is not to contrast courses at the same level, the remainder of the analysis will aggregate all introductory and upper-division students separately.

The breakdown of overall participation rates is given in Table I. These rates are somewhat higher than what has been observed for post-test participation in either paper-based or online RBAs in previous studies. For example, the LASSO study found for their introductory population an average post-test response rate of 66% for paper-based RBAs and 50% for online RBAs [13,14]. We also have historical participation rates available for three of the upper-division courses and both of the two introductory courses in the dataset. Average historical participation rates for these courses were between 60% and 85% for both the upper-division courses and introductory courses. Response rates in paper administrations were dictated by attendance during one lecture or tutorial session during the last week of classes. All of these courses include interactive elements (e.g., clicker questions) for which students receive regular course or extra credit, but attendance was not otherwise required. These same courses saw participation rates between 79% and 97% in the current dataset, suggesting that the participation rate for these courses actually increased somewhat when the RBA was given online. The fact that the instructors in our dataset offered a meaningful amount of regular course credit for students who participated in the RBA, independent of their performance, likely contributed significantly to this increase in participation. We do not have consistent access to data on the racial or gender distributions for the students in our dataset and thus do not report this breakdown here.

TABLE I. Overall participation rates for the introductory and upper-division populations. Participation rates (Rate) represent the percentage of the total course enrollment for which we collected valid responses (N_{valid}) in the final dataset. Individual course participation rates varied from 70% to 100%.

	N_{valid}	N_{roster}	Rate
Introductory	1287	1543	83.4%
Upper division	308	336	90.2%

On the first page of the assessment, students were instructed to complete the RBA in one sitting without the use of outside resources such as notes, textbooks, or Google. To capture students' online behaviors, we embedded JavaScript code into the online prompts to look for instances of students copying text, printing from their browser, or clicking into another browser window. For the upper-division population, the code only recorded when a student copied text, but did not record what text was copied. However, for the introductory population, we also collected data on what question the text was copied from. In all cases, these behaviors were time stamped to determine when each action occurred and how many times each student exhibited that behavior. This JavaScript code could only detect activities that happen at the browser level; activities at the computer level (e.g., taking a screenshot or clicking into another program) were not recorded by the code. While such data would be useful, modern browsers nearly all have security features to prevent cookies and scripts in browsers from collecting information on activities happening outside the current browser window.

For browser print commands (e.g., "control-p") and copy text commands (e.g., "control-c"), the primary data collected were when and how often these commands were

issued. Data on browser focus were somewhat more complex. The code was designed to listen for a change in browser focus, and then record whether the RBA tab was visible 4 sec after the browser focus event occurred. This allows for a variety of patterns in focus data as students click in and out of the browser tab, which they sometimes did rapidly and repeatedly. However, in general, if a student clicks into a new browser tab and stays in that tab for more than 4 sec, the code would record a single browser focus event and tag it "hidden." A hidden browser focus event most often means that the student left the RBA without returning to it within 4 sec. Alternatively, if the student clicked into another browser tab and then clicked back into the RBA within 4 sec (and remained there for more than 4 sec), the code would record two browser focus events—one for the click out and one for the click in—and would tag both as "visible." A single "visible" browser focus event most often means that the student returned to the RBA for more than 4 sec after having left it for any amount of time.

In addition to the data on students' online behaviors, we collected students' scores on the assessment in order to compare the prevalence of students' behaviors with their performance on the RBA. Total time spent on the assessment was approximated using the time elapsed between the time the student clicked on the link for the RBA and the time they submitted it. As we discuss in more detail in the next section, this duration is only approximate because it does not account for time the student might have spent away from the RBA (e.g., in another browser tab, or not on their computer).

IV. RESULTS

Here, we examine data on students' behaviors on online RBAs to determine how prevalent specific online behaviors were for this population of students. We also examine correlations between these behaviors and students' performance on the assessments overall.

A. Print events

The primary concern associated with students printing or saving RBAs is that these students might publicly post the assessment and thus breach the security of the assessment by making it available to other students. Because the online RBAs were designed to mirror the paper-based versions, each had 10–15 individual pages that the students would work through to see all questions. This means that to present a significant threat to the security of the assessment, a student would need to print each page of the assessment separately, and in so doing, would register multiple print commands. To determine the prevalence of students printing their browser page, we include responses from the full dataset, including responses marked as "invalid" from, for example, students who did not ultimately submit the RBA. In this full dataset of 1879 student responses, only five (two from the introductory population and three from the upper division) had recorded print events. Of these, 3 students, all from the upper-division population, had multiple print commands consistent with having saved all or the majority of the assessment pages. The remaining 2, both from the introductory population, had only 1–2 distinct print events meaning they could, at most, have saved only a small number of questions. It could be that after beginning the process of saving the questions, these students realized the process would require saving each page of the assessment individually and gave up.

Print commands themselves do not necessarily indicate a student who is intending to breach the security of the assessment. In fact, one of the instructors (S. J. P.) reported interacting with a student during help hours in which the students pulled up screenshots of the assessment which he had taken to study from after the fact. The student made no attempt to hide the screenshots and was upfront with his motivation for taking the screenshots as a study tool. Moreover, even if a student did post the RBA prompts online, without corresponding solutions, which were never released to the students, it is not clear that access to the RBA prompts alone actually represents a significant threat to the assessment's security or validity. Additionally, as is standard for paper-based assessments, the formal names or acronyms for the assessments (e.g., the CUE) were not provided to students in the online versions.

To test for any immediate security breaches of the assessments, we Googled the prompts for each question on all four RBAs used in this study several weeks after the assessments had closed. The results of these searches varied significantly for the introductory and upper-division assessments. For the two upper-division assessments (the CUE and QMCA), there was no indication that the item prompts or their solutions had been uploaded in a way that ranked high in Google's listing. However, as Google's algorithm can change based on search patterns, it is likely necessary to do this type of verification periodically to ensure no solutions have surfaced. In several cases, Googling the item prompts pulled up PER publications on the test itself, and some of these publications included supplemental material which contained the grading rubrics for the assessment in one form or another (open ended or multiple choice). It is worth noting that in all cases these rubrics were buried at the end of a long publication or thesis and not clearly marked, and it is not clear if a student who was unfamiliar with the specific publications (or the nature of academic publication more generally) would be able to locate the rubrics without considerable persistence. However, this suggests that the greatest threat to the security of the upper-division RBAs in an online format may actually be our own publications combined with the fact that the premier PER publication venue is open access.

Attempts to Google the prompts to the two introductory RBAs (the FMCE and BEMA), however, yielded very

different results. Searching prompts for items on these assessments pulls up images of the exact prompts from the assessment, and accompanying solutions are available on paid solution sites like Chegg or Course Hero. Any student with an existing subscription to these sites would like be able to find solutions to the FMCE or BEMA questions with relative ease. These solutions predate this study, and thus represent breaches of security that occurred previously. The larger online presence of both the introductory RBA prompts and solutions has at least two possible contributing factors. First, introductory (and largely nonphysics major) students may be more likely to engage in behaviors that facilitate quick completion of online assignments rather than prioritizing deep learning of the material. Thus, they may be more likely to look for, and share, course materials online. Second, both the FMCE and BEMA are considerably older and more extensively used than the CUE and QMCA. It may be that solutions to any RBA will eventually make their way online given sufficient time and use, and that the CUE and QMCA are not old enough or common enough to have achieved a significant online presences. It is also possible that even in the absence of the presence of solutions, resources related to introductory physics content may be more common and easily located online than resources related to the more advanced upper-division content. We will discuss additional implications of these patterns in Sec. IV C.

B. Browser focus events

Online RBAs introduce a potential for students to become disengaged from the assessment in a way that is less likely in paper-based administrations. Loss of browser focus is one proxy for students disengaging from the RBA. Focus events were the most common events in the dataset with roughly half of the students (46%, $N = 562$ of 1287 in introductory; 52%, $N = 159$ of 308 in upper division) with at least one browser focus event in which their RBA window became hidden for more than 4 sec. For these students, we examined trends in the number and duration of browser focus events by grouping them to isolate sustained changes in browser visibility. In other words, if a students' survey page becomes hidden, how long before it becomes visible again, independent of whether there are additional browser hidden events in between (indicating that the student clicked back into the survey window, but did not remain there for more than 4 sec)? Here, we will report median and max duration, as the presence of even a small number of outliers makes the average less meaningful.

Table II reports information on the number and duration of browser focus events in the dataset. While Table II reports data for the introductory and upper-division students separately, the trends are comparable between the two levels. These trends suggest that a large fraction of students in the dataset did click out of the assessment tab one or more times while taking the RBA; however, roughly two-thirds of the time they were away from the RBA for no more than 1 min and less than 10% left the assessment for longer than 5 min. Moreover, just over one-third of students left the assessment only once. Here, we have selected 1 min as a relevant time frame because, in our experience implementing assessments like these in in-class environments, this time frame is generally comparable to how long a student might "space out" while taking the RBA during class. However, it is not possible for us to know what the students were doing while their browser was hidden; thus, we cannot determine if the loss of focus was distraction related or related to use of internet resources to improve their performance on the assessment.

To investigate the impact of browser loss of focus, we examined whether the appearance or duration of loss of focus events correlated with students' scores on the assessment. In as much as browser loss of focus could be a proxy for distraction, it might be guessed that students with loss of focus events would score lower than others on the RBA. Alternatively, if the loss of focus is associated with use of internet resources (see Sec. IV C), we would anticipate students with loss of focus events to potentially score higher. To account for difference in average score between courses in the dataset, z scores calculated relative to the average score for each individual class were used in calculating correlations. Students with loss of focus events scored higher on average by roughly a quarter of a standard

TABLE II. Duration and number of sustained browser hidden events in the introductory and upper-division student population. For reference, the total number of valid responses in the introductory and upper-division datasets was $N = 1287$ and $N = 308$, respectively.

	Introductory	Upper division
Total number of students with 1 or more focus event	562	159
Number of focus events per student	Median—2 (Max—43)	Median—2 (Max—59)
Number of students with only 1 focus event	219	66
Number of students with 10 or more focus events	91	20
Total number of focus events	2860	725
Duration of focus events	Median—21 sec (Max—66.7 hr)	Median—34 sec (Max—29.3 hr)
Number of focus events less than 1 min	2264	479
Number of focus events greater than 5 min	149	70

deviation than other students for the introductory RBAs (i.e., a z-score difference of 0.26) and lower on average by roughly one-fifth of a standard deviation for the upper-division RBAs (i.e., a z-score difference of −0.19). The difference in performance was statistically significant in the case of the introductory courses (Mann-Whitney U $p = 0.001$) though small (Cohen's $d = 0.26$), and was not statistically significant for the upper-division population. Additionally, we examined the Spearman correlation coefficient between the total time students spent with their browser hidden relative to their score on the assessment. We selected the Spearman correlation because it is less sensitive to the presence of outliers than the other coefficients. Consistent with the differences in average score, we found a statistically significant, though small, correlation between score and total time away from the assessment tab for introductory students ($r = 0.16,\ p = 0.0001$) and no significant correlation for the upper-division students ($r = -0.1,\ p = 0.2$).

C. Copy events

The primary concern associated with students copying text from an online RBA is that students may do so in order to search the internet in an attempt to "look up" the correct answer. Table III shows the prevalence of copy events within our dataset, showing that roughly one-tenth of the students in the dataset had one or more copy events. A copy event, on its own, does not necessarily mean that the student was attempting to web search answers to the questions. However, if a student copies text with the intention of searching the web for that text, this behavior would most likely be characterized by a copy event followed immediately by a sustained browser hidden focus event. To investigate this, we looked for copy events followed within 5 sec by a sustained browser loss of focus event and counted how many times this occurred for each student. We found that more than three-quarters of the copy events ($N = 654$ of 861 events for introductory, and $N = 56$ of 67 events for upper division) fell into this category. This indicates that a majority of copy events were immediately followed by the student switching into a new browser window and remaining there for more than 4 sec, consistent with the pattern we would expect if they were trying to web search the item prompts. The remaining copy events that were not followed by a loss of focus event were typically characterized by either the first of two quick consecutive copy events followed by a single loss of focus event or single copy events not connected temporally with a loss of focus event.

Given this pattern, we also examined whether the students with copy events had any difference in performance from other students. For the introductory RBAs, students with copy events scored higher than students without copy events (average z-score difference of 0.45). This trend was exactly flipped for the upper-division RBAs where students with copy events scored lower (average z-score difference of −0.46). This difference was statistically significant in both cases (Mann-Whitney U $p < 0.001$) and of moderate effect size (Cohen's $|d| = 0.46$ in both cases).

In the second semester of data collection, in which all data from the introductory RBAs were collected, additional Javascript code was included; this code collected information not only on when students copied text, but also from which question prompt they copied that text. We used this information to determine if a student who copied the text of an item was more likely than the rest of the students to get that specific question correct. To determine this, we looked at each question individually and counted how often a student who copied text from that question got it correct vs got it incorrect. Similarly, we counted how often students who had not copied text from that question got it correct vs incorrect. The result was a 2×2 contingency table with columns denoting whether or not the student copied text from that question and rows denoting whether the student got the question correct or not. We then summed the tables across all questions and the resulting contingency table is given in Table IV.

Table IV shows that, on average, when introductory students copied text from an item they responded correctly to that item 77% of the time. Alternatively, introductory students who did not copy text from a particular item responded correctly to that item only 58% of the time, on average. This difference in frequency is statistically significant (Chi-squared $p \ll 0.001$). This shows that

TABLE III. Number of copy events detected in the introductory and upper-division populations. For reference, the total number of valid responses in the introductory and upper-division datasets was $N = 1287$ and $N = 308$, respectively. The CUE, QMCA, FMCE, and BEMA have 16, 38, 47, and 31 questions, respectively.

	Introductory	Upper division
Number of students with copy events	147	22
Median number of copy events per student	4	2
Max number of copy events per student	54	11
Total number of copy events	861	67

TABLE IV. Contingency table breaking down how often students responded to a question correctly relative to whether they had copied text from that question. This table includes data from all questions; thus, each count in the table represents a response from one student to one question. Percentages are given with respect to the total number of copy or noncopy events.

	Copied text	Did not copy text
Correct response	559 (77%)	28 291 (58%)
Incorrect response	163 (23%)	20 596 (42%)

students who copied text from a question were more likely to get that question correct than students who did not. We can also look at whether a student who copies text from one or more questions scores higher, on average, on those questions than on the subset of questions from which they did not copy text. To determine this, we focused just on the $N = 147$ introductory students who had one or more copy events. We then calculated z scores for their performance on the subset of questions where they copied text and z scores for their performance on the subset of questions where they did not copy text. We then average the two resulting scores across all students to determine whether students perform better on average on questions where they copied text relative to questions where they did not. We found that the z score on the subset of copied questions was higher on average by just under half a standard deviation (i.e., a z-score difference of 0.44). This difference is statistically significant (Mann-Whitney U $p \ll 0.001$) and of moderate size (Cohen's $d = 0.4$).

Together, these results suggest that, in the introductory courses, roughly 10% of students do try to look up the answers to the RBA and that doing so appears to improve their performance. In a high-stakes testing environment, this trend would be extremely problematic as it would imply that an individual students' score could not be reliably interpreted. However, RBAs within PER are intended to be low-stakes measures of group (rather than individual) performance; it is widely considered inappropriate, for a range of theoretical and practical reasons, to use RBAs as a measure of individual student performance [25]. So the question then becomes, what impact does this copying behavior have on the average score for the class as a whole and, thus, our ability to interpret and compare across online and paper-based administrations of the RBA. To determine this, we compare the average score for the full introductory dataset relative to the overall score for just the subset with no copy events. We examine this for both the total score and the scores for individual items.

Removing all students who had copy events from the introductory dataset resulted in a drop in overall average score of roughly 1.1%. This difference represents a very small effect (Cohen's $d = 0.05$) and was not statistically significant (Mann-Whitney U $p = 0.2$). To get a "worst-case-scenario" sense of the size of this effect, we also calculated the average score for the class assuming that on any question where a person copied text they would have gotten that question wrong otherwise. In other words, we zeroed out the score for any question where a student copied text and used this to calculate their "worst-case" score. We then calculated the course average now including all students but with the copied questions zeroed out. The difference in course averages rose to roughly 1.2% in this worst-case-scenario. Together, these results suggests that looking up answers to the RBA online, while certainly significant in its impact on individual student's scores, had a statistically and practically negligible impact on the overall course average. By individual item, the difference in average item score generated by removing students who copied that item had a range of [−0.27%, +1.4%] with a mean of 0.29%, suggesting that the impact of students looking up answers on individual average item scores is, in practice, negligibly small. To be clear, this analysis focuses exclusively on the impact of copying behavior on the overall course averages, and does not suggest that the copying behaviors had no impact on *individual* students' scores; in fact, the analysis earlier in this section showed that it does. The statistically and practically insignificant impact for the *class average* is linked to the fact that only a small number of students exhibited copying behaviors and typically only on a small number of questions (see Sec. V for additional discussion).

D. Time to completion

We also examine the total amount of time to completion for each student to determine whether student's scores are related to how long it took them to complete the assessment. Total time data are calculated by comparing the recorded time when the student first opened the survey link to when they made their final submission of the survey. This does not remove periods when browser focus was lost, and can even include a period when the survey window was closed and later reopened. As such, these duration do not necessarily reflect the amount of time a student actually worked on the assessment, merely the amount of time that passed between them opening and submitting the assessment. For the majority of students (65%, $N = 843$ of 1287 in introductory; 78%, $N = 239$ of 308 in upper division), the total time between start and submit fell within a time frame of 15 and 60 min, consistent with what would be required of a student taking the RBA in class. Total time spent on the RBA showed a significant (though small) correlation with z score on the assessment only for the introductory students (Spearman $r = 0.3$, $p \ll 0.001$).

We can also use the focus data to modify the raw time data by subtracting out the total time for each student during which their survey window was hidden, suggesting they may not have been working on the assessment. Doing so does not significantly shift either the number of students whose total time (now excluding time away from the browser) falls between 15 and 60 min or the correlation of time with score on the assessment for either introductory or upper-division students.

V. DISCUSSION AND LIMITATIONS

We collected online responses to four research-based assessments spanning both introductory and upper-division content. This work is part of ongoing research to determine whether students' performance on RBAs shifts when these assessments are given online. For three of the courses in the

dataset (1 introductory and 2 upper division), we also have historical scores from students in these same classes with the same instructor where the RBA was given on paper and during class. Comparisons of the online and in-class scores showed the online scores being roughly 5% lower. This difference was statistically significant only in the case of the introductory population (two-tailed t test, $p = 0.001$) though the effect was small (Cohen's $d = 0.13$). The decrease in average score appeared to be largely driven by the presence of a larger tail of the distribution (in terms of grades) in the online administrations. This, combined with the higher participation rates in the online administrations (see Sec. III) suggests that administering RBAs online to an upper-division population encourages more of the lower performing students to participate.

In addition to students' responses to the RBAs, we also collected data using embedded JavaScript code on students' online behaviors such as copying text, printing browser pages, and losing browser focus by clicking into other browser tabs. We found that only a small number of students (less than 0.5%) printed or copied item prompts in a manner that suggested they were attempting to save some or all of the item prompts. Such behavior primarily represents a potential concern with respect to test security if students chose to post the assessment prompts online. However, we have anecdotal evidence that at least some of these students were saving the prompts solely for their own future studying and with no intention of sharing them. How much of a concern maintaining test security is may also vary between introductory and upper-division RBAs. Our own attempts to look up solutions to the RBAs used in this study showed that item prompts and solutions to the FMCE and BEMA are already available online on paid solution sites. Alternatively, we found no evidence of item prompts or solutions for the CUE and QMCA. Thus, test security has already been at least partially breached for the introductory RBAs, but appears to be largely intact for the upper-division RBAs. This may be a reflection of the fact that the FMCE and BEMA are both older and more widely used assessments than the CUE and QMCA.

We also collected data on how often and for how long students clicked out of their RBA browser tab as a proxy for distraction. Such behavior was common, with roughly half the students engaging in online behaviors resulting in loss of browser focus and indicating that the students may have disengaged from the RBA for a period of time. However, roughly two-thirds of the periods where students lost browser focus lasted less than 1 min, and less than 10% of the periods lasted for longer than 5 min. The total amount of time spent away from the assessment tab had a small but statistically significant positive correlation with overall score on the RBA only for the introductory population. Thus, we argue that while the potential for distraction and disengagement certainly increases with online RBAs, our data suggest the majority of students do not become disengaged for long periods and that this disengagement does not appear to negatively impact their performance. The slight positive correlation that appears for the introductory students is unexpected when considering time away from the assessments tab as a proxy for disengagement. This trend may be driven by students who navigated away from their RBA browser tab when accessing internet resources to assist them in completing the assessment.

Evidence of copying text was observed in roughly 10% of the students in our sample. Roughly three-quarters of these copy events were immediately followed by a browser focus event in which the RBA tab became hidden. Such a pattern is consistent with what we would observe if students were attempting to Google the item prompts in an attempt to determine the correct answers. While it is not possible for us to determine for certain if that is what the students were doing, the pattern is suggestive. Moreover, students with copy events had statistically different score distributions than the rest of the populations. However, the trend differs between the introductory and upper-division students. Upper-division students with copy events scored lower than other students, while introductory students scored higher. Using information on which specific questions students copied text from, we found that students who copied text for a particular question more often got that question correct. Moreover, we found that students scored, on average, higher on the subset of items from which they copied text than from the items they did not.

Taken together, the findings summarized above are consistent with the follow interpretation. A small subset of students do attempt to Google item prompts when taking online RBAs, and, as evidenced by the lower performance of these students in the upper-division population, these students may deferentially include lower performing students. In cases where the solutions to the specific RBAs are not easily accessible online (e.g., the CUE and QMCA), copying and Googling text does not improve students' scores. Alternatively, in cases where the solutions to the specific RBA are available online, copying and Googling text does result in an improvement in students' performance. However, because the improvement to students' scores is, on average, small (roughly one-third of a standard deviation of improvement in average score) and impacts only a small fraction of students, the impact of this behavior on class average scores overall or by question was negligibly small.

Overall, our findings suggest that, while students in our sample engaged in a variety of online behaviors, none of these behaviors resulted in a change in the population's average performance that would threaten our ability to interpret this performance or compare it to paper-based implementations of the RBA. However, this held largely because only a small component of the student population actually engaged in some of these behaviors. Should the

number of students engaging in, specifically, copying behaviors increase in the future, the impact of these behaviors on the course average may increase.

The only effect observed in the current study that presents a concern for comparisons of online and paper-based RBAs was a consistent roughly 5% drop in overall average score relative to historical paper-based implementations in these courses. This drop may be a result of the larger participation rates observed in the online administrations and, thus, the inclusion of a larger component of the lower performing tail of the student population. Rather than being a problem, we argue this actually represents an advantage for the online RBAs in that they appear to provide a broader sample of the student population.

The work presented here has some important limitations. The code that captured students' online behaviors can only detect actions at the browser level, meaning that actions at the computer level (like switching into a new program) cannot be detected. For this reason, our data should be interpreted as a lower bound on the appearance of these behaviors. Replication of this work with additional RBAs, with a broader student population, and in future semesters will be important to ensuring that these results hold across different tests, a broad student population, and time. However, these results do suggest that online assessment is a promising alternative that brings with it many potential logistical advantages.

ACKNOWLEDGMENTS

This work was funded by the CU Physics Department. Special thank you to the faculty and students who participated in the study and the members of PER@C for all their feedback.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter rbai-1: research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[2] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. Deardorff, R. J. Allain, S. W. Bonham, M. H. Dancy, and J. S. Risley, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, Research-based Reform Univ. Phys. **1**, 2 (2007).

[3] D. Hestenes, Toward a modeling theory of physics instruction, Am. J. Phys. **55**, 440 (1987).

[4] C. H. Crouch and E. Mazur, Peer Instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[5] B. R. Wilcox, B. M. Zwickl, R. D. Hobbs, J. M. Aiken, N. M. Welch, and H. J. Lewandowski, Alternative model for administration and analysis of research-based assessments, Phys. Rev. Phys. Educ. Res. **12**, 010139 (2016).

[6] D. MacIsaac, R. P. Cole, D. M. Cole, L. McCullough, and J. Maxka, Standardized testing in physics via the world wide web, Electron. J. Sci. Educ. **6**, 1 (2002).

[7] B. Van Dusen, L. Langdon, and V. Otero, Learning assistant supported student outcomes (lasso) study initial findings, in *Proceedings of the 2015 Physics Education Research Conference, College Park, MD*(AIP, New York, 2015), p. 343.

[8] https://www.physport.org/assessments/, 2015.

[9] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[10] B. R. Wilcox, M. D. Caballero, C. Baily, H. Sadaghiani, S. V. Chasteen, Q. X. Ryan, and S. J. Pollock, Development and uses of upper-division conceptual assessments, Phys. Rev. ST Phys. Educ. Res. **11**, 020115 (2015).

[11] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[12] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[13] M. Jariwala, J. Nissen, X. Herrera, E. Close, and B. Van Dusen, Participation rates of in-class vs. online administration of low-stakes research-based assessments, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2017), p. 196.

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

[15] J. Nissen, M. Jariwala, X. Herrera, E. Close, and B. Van Dusen, Performance on in-class vs. online administration of concept inventories, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2017), p. 272.

[16] D. Zandvliet and P. Farragher, A comparison of computer-administered and written tests, J. Res. Computing Educ. **29**, 423 (1997).

[17] R. K. Ladyshewsky, Post-graduate student performance in supervised in-class vs. unsupervised online multiple choice tests: Implications for cheating and test security, Assessment and Evaluation in Higher Education **40**, 883 (2015).

[18] A. C. Bugbee, Jr., The equivalence of paper-and-pencil and computer-based testing, J. Res. Computing Educ. **28**, 282 (1996).

[19] W. M. Haney and M. J. Clarke, Cheating on tests: Prevalence, detection, and implications for online testing, in *Psychology of Academic Cheating* (Elsevier, New York, 2007), p. 255.

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

[21] H. R. Sadaghiani and S. J. Pollock, Quantum mechanics concept assessment: Development and validation study, Phys. Rev. ST Phys. Educ. Res. **11**, 010110 (2015).

[22] B. R. Wilcox and S. J. Pollock, Validation and analysis of the coupled multiple response Colorado Upper-Division Electrostatics Diagnostic, Phys. Rev. ST Phys. Educ. Res. **11**, 020130 (2015).

[23] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[24] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[25] P. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, in *Getting Started in PER*, Vol. 2 (American Association of Physics Teachers, College Park, 2009).