Abstract

The COVID-19 pandemic posed many disruptions to higher education assessment in 2020. At James Madison University (JMU), ensuing modifications to long-standing, university-wide assessment necessitated unproctored remote testing instead of the typically proctored, onsite assessment. Applying such modifications to low-stakes educational assessment raises validity concerns. JMU's assessment model allowed us to explore the effect of the different test administrations, taking into account pre-existing trends in cohorts' performance. We compared assessment results on three tests (history, global issues, and scientific reasoning) between the 2020 entering class (tested remotely) and the previous four cohorts (tested in-person). Our results revealed lower test performance and a bimodal distribution of effort scores in students tested remotely in 2020, but only on the more cognitively demanding scientific reasoning test, compared to the less arduous tests, history and global issues. Implications and limitations are discussed.

AUTHORS

Sarah Alahmadi, M.S.
James Madison University

Christine E. DeMars, Ph.D.
James Madison University

Large-Scale Assessment During a Pandemic: Results from James Madison University's Remote Assessment Day

Assessment efforts in higher education were among the many domains and practices that COVID-19 has disrupted in 2020. A report published by the National Institute for Learning Outcomes Assessment (NILOA) revealed that 97% of 813 higher education professionals who held assessment-related roles indicated that changes to their assessment were necessitated in response to COVID-19, especially with regards to modifying assignments or assessments (Jankowski, 2020). At James Madison University (JMU), assessment modifications were required not only at the course- and program-level, but also at the university-level. For more than 30 years, JMU has been collecting longitudinal data assessing learning outcomes for every cohort. Students are assessed twice, first as incoming first-year students (i.e., before completing any classes) and again after completing 45-70 credit hours. Such a model allows for a longitudinal assessment of learning growth. Additionally, having assessed students for the last 30 years allows us to observe larger trends in learning improvement across cohorts.

JMU's Assessment Day model and its logistics were comprehensively described by Pastor et al. (2019). The Assessment Days typically involve around 4,000 students tested in one of three proctored, 2-hour sessions. Different groups of incoming students are randomly assigned different configurations of Assessment Day instruments. Some assessments are completed using paper-and-pencil while others are computer-based. Proctors play an

CORRESPONDENCE

Email
alahmasi@jmu.edu

important role on Assessment Days as they ensure that tests are completed properly, noises are minimized, and students are motivated and aware of the importance of the Assessment Day. However, changes were necessary for the 2020-2021 academic year: Assessment was conducted remotely due to the COVID-19 pandemic, whereas all assessment was conducted in person in previous years.

Conducting a remote Assessment Day constituted many modifications to the abovementioned procedures (Pastor & Love, 2020). Instead of being tested on a specific day, students were allowed a three-week[1] window to complete the assessments via the links they received by email. The format of testing changed from paper-and-pencil to computer-based. Participation rates were somewhat lower. These changes raised several questions: Do students tested remotely in Fall 2020 score comparably to students tested in-person in the previous cohorts? If there are differences, are the differences similar across different tests? Also, do students tested remotely in Fall 2020 report test-taking effort similar to the effort reported by students tested in-person in the previous cohorts?

These findings indicate that the results of low-stakes testing may not precisely reflect individual differences in proficiency; rather the results are confounded by other factors, such as motivation or effort, rendering the validity of such results questionable.

Assessment Day testing is considered low-stakes testing, because students' performance bears no direct personal consequences. Thus, students could vary in the amount of effort they expend on assessment tests. Low effort has been found to affect performance by producing scores that underestimate ability (Wise & DeMars, 2005). In their review of examinee effort in low-stakes testing, Wise and DeMars computed differences between groups tested under motivating and less motivating conditions. Across 12 studies, they found that, on average, students tested under more motivating conditions performed more than one-half standard deviation higher. These findings indicate that the results of low-stakes testing may not precisely reflect individual differences in proficiency; rather the results are confounded by other factors, such as motivation or effort, rendering the validity of such results questionable. There are several strategies that could be employed to improve students' motivation, such as increasing the stakes of testing and selecting less cognitively taxing test designs. JMU utilizes both strategies by (1) making Assessment Days *semi*-consequential by not allowing students to register for future semesters if they did not attend Assessment Day, and (2) devising tests that contain mainly multiple-choice questions as opposed to essay questions; a strategy that has been shown to be less cognitively-overwhelming, maintaining higher levels of effort from students (DeMars, 2000). Also, students are made aware of the importance and value of Assessment Day before they complete their tests. In a typical year, proctors would ensure that students completed all the tests within the allotted time and that no students left the testing room early.

Moving Assessment Day online in Fall of 2020 raised several validity concerns that often accompany low-stakes, unproctored internet test (UIT) administrations. In general, implementing a UIT design entails unstandardized testing conditions among examinees regarding, to name a few, the amount of time spent completing the tests, environmental noise, and technological issues. While—specific to our interest—the fact that the test is *low-stakes* alleviates the usual UIT concerns around examinee cheating, it brings about questions related to examinee motivation and effort. Empirical evidence is mixed with regards to whether low-stakes UIT produces differences across test scores by introducing construct-irrelevant variance. One study that compared examinee performance in proctored versus unproctored online settings found no significant differences (Hollister & Berenson, 2009). Another study that examined the effect of *web-based* tests in several conditions—including proctored, in-person and unproctored, remote—reported no differences (Templar & Lange, 2008). Conversely, there is some evidence suggesting higher performance in web-based, remote unproctored cognitive tests (Karim, Kaminsky, & Behrend, 2014).

These findings collectively provide some evidence that differences in performance may occur. However, one study that looked specifically at performance differences between low-stakes online-proctored tests and online-unproctored tests found some reassuring results (Rios & Liu, 2017). The study examined differential performance and test-taking behavior based on whether online tests were proctored. Test-taking behavior was examined

[1] The window was later extended due to disruptions in on-campus courses early in the semester.

via keystroke data (the frequency of item views, items omitted, and items not-reached), and response time data (total testing time and rapid-guessing time). The results showed negligible and insignificant differences in terms of test-taking behavior as well as test scores between those whose online test was proctored and those whose online test was unproctored. These findings suggest that in low-stakes online testing, there are no meaningful implications for the absence of proctoring.

The question remains whether administering low-stakes tests remotely versus in-person would have differential implications for assessments. There is not yet any individual study that compares performance differences among college students on cognitive *low-stakes* tests in an in-person proctored, paper-and-pencil administration versus an online unproctored administration. By sharing the results of our remote Assessment Day, we hope to shed some light on this unexplored area. In this paper, we compare the scores from several tests delivered remotely in 2020 to the scores from the same tests administered in person in previous years, to see if there are performance differences and if those differences vary by test. We then examine differences in self-reported effort and in time spent testing as possible explanations of differences in test performance.

Method

Participants

Participants were first-year students entering the university in 2016-2020. All students were required to participate in Assessment Day, but different students were randomly assigned to each assessment instrument. For this study, data were used from all students who consented to having their results used for research and completed one of the three selected instruments, described below. Demographic information about the participants is shown in Table 1. In 2016-2019, students who did not complete their assessments were prevented from registering for the next semester until they participated in a make-up session. In 2020, there were no consequences for not participating. As described earlier, the assessments in 2016-2019 were completed at an assigned time, on paper, in a group setting, supervised by a proctor, whereas the 2020 assessments were completed anytime within a 3-week window, on computer, in a setting of the student's choice (generally home or dorm room), and unproctored.

There is not yet any individual study that compares performance differences among college students on cognitive lowstakes tests in an in-person proctored, paper-and-pencil administration versus an online unproctored administration. By sharing the results of our remote Assessment Day, we hope to shed some light on this unexplored area.

Assessment Instruments

Three of the General Education assessments were chosen for this study because they have been administered for at least five years and thus have a history from which to judge whether scores in 2020 were within the range of year to year fluctuation or represented a departure from past trends. These assessments span different subject areas and test lengths. The selected instruments were developed by faculty to target students' knowledge in history, global issues, and scientific reasoning. We also administered an assessment of test-taking motivation and effort, the Student Opinion Survey (Sundre & Moore, 2002).

Knowledge of history and political science is assessed using a 40-item test, with a possible number correct score range of 0 to 40. Knowledge of global issues is assessed by 31 items, with a possible number correct score range of 0 to 31. Scientific reasoning is assessed by 66 items, with a possible range for number correct score between 0 and 66. Lastly, effort and motivation are measured by a 5-item survey. Students indicate their agreement level with statements regarding how much effort they expended on a 5-point Likert scale ranging from 1 = *Strongly Disagree* to 5 = *Strongly Agree*. The possible total score range is 1 to 5 after taking the average over the five items.

Results

Test Scores

To make comparisons of cohort performance across the differently scaled assessments, we standardized the scores. The standardization was based on students with no course credit tested in 2016-2019; for these students, the mean was set to zero and

Table 1
Participants

Year	Test	*N*	% Female	% In-State Residents	% non-Hispanic White
2016	History	1041	60%	75%	76%
	Global Issues	821	61%	73%	77%
	Scientific Reasoning	817	61%	73%	78%
2017	History	996	59%	73%	79%
	Global Issues	1148	60%	74%	76%
	Scientific Reasoning	734	58%	73%	75%
2018	History	1027	58%	73%	77%
	Global Issues	767	60%	73%	75%
	Scientific Reasoning	745	62%	69%	78%
2019	History	1178	58%	77%	77%
	Global Issues	1031	60%	75%	76%
	Scientific Reasoning	458	60%	74%	77%
2020	History	841	62%	75%	76%
	Global Issues	840	63%	77%	77%
	Scientific Reasoning	457	67%	76%	77%

We observe a pattern of decreasing scores over the years on history and global issues, but fluctuating scores on the scientific reasoning test, with a large drop in 2020. In the scientific reasoning assessments, however, there was not a clear trend prior to 2020, and the 2020 group exhibited a more drastic decrease.

the within-group pooled-standard deviation was set to one. See Figure 1 for standardized mean comparisons across the last five cohorts for only those with no credits. Because the within-group standard deviation was set to one, differences in Figure 1 can be interpreted similarly to Cohen's *d*. We observe a pattern of decreasing scores over the years on history and global issues, but fluctuating scores on the scientific reasoning test, with a large drop in 2020. It seems that students in 2020 conformed to the general pattern of slightly decreasing scores year by year on the history and global issues assessments. The linear trend was statistically significant (history: $F_{1, 5078} = 21.35$, $p < .001$; global issues: $F_{1, 4602} = 47.70$, $p < .001$), but there were no significant differences among the cohorts beyond the linear trend (history: $F_{3, 5078} = 2.91$, $p = .4677$; global issues: $F_{3, 4602} = 188$, $p = .1303$).[2] In the scientific reasoning assessments, however, there was not a clear trend prior to 2020, and the 2020 group exhibited a more drastic decrease. A contrast between 2020 and the mean of the previous years showed that 2020 scores were significantly different ($F_{1, 3206} = 180.63$, $p < .001$). The 2020 mean was 0.75 standard deviations below the mean for the previous years.

Additional information about student test performance can be gained by examining the distribution of scores. In Figure 2, the score distribution for scientific reasoning did not just shift lower—the shape of the distribution changed. The mode of the distribution in 2020 was located just below the mode of previous cohorts, but there was a secondary mode of lower scores. A substantial portion of the students scored much lower than previous cohorts.

[2] There were five groups, so the omnibus F-test was partitioned into a 1-*df* linear trend a 3-*df* test of the remaining variance. The latter test was of interest in this study, and answered the question: Beyond the linear trend, were there any significant differences in the group means?

Figure 1
Mean Standardized Scores across Cohorts

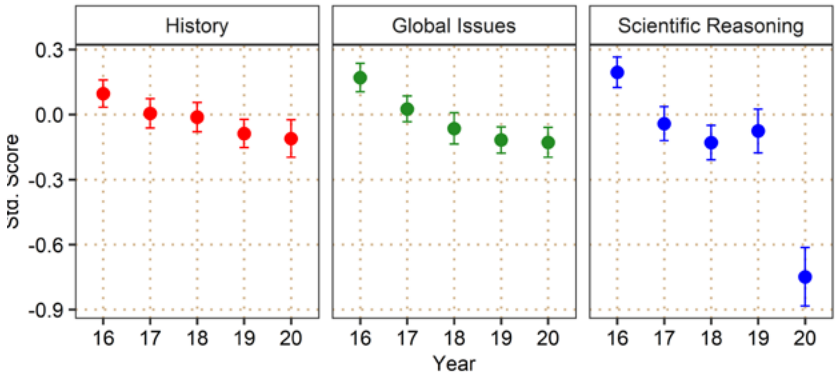
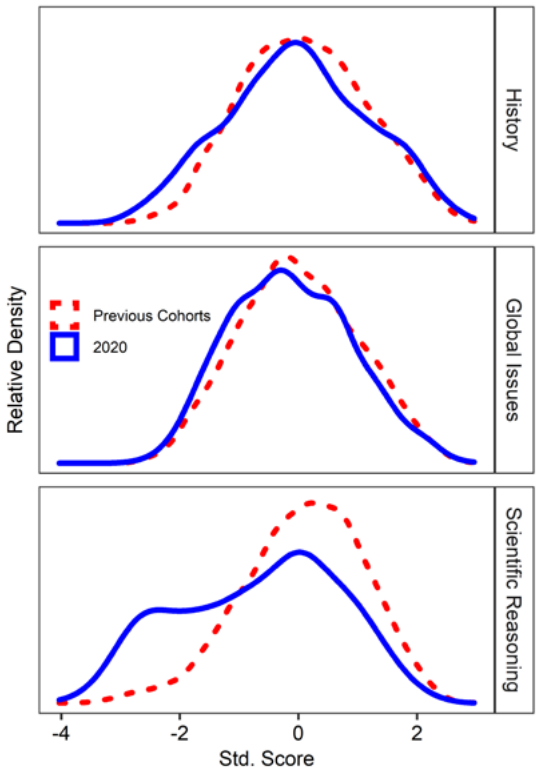


Figure 2
Distribution of Test Scores



Students enter the university with varying levels of course credit, such as transfer or AP credit. The comparison in Figure 1 used data only from students with no course credit, to avoid the possibility of confounding administration conditions with differences in the proportion of students with course credit. However, the performance of students with course credit may also be of interest. Table 2 presents raw mean scores on the three tests assessing students in history, global issues, and scientific reasoning, overall and broken down by course credit. For simplicity, we report mean scores for this year's cohort, 2020, and the previous four cohorts (from 2016 to 2019) combined. Students in the "No credit" column and the 2016-2019 row were used for setting the standardized metric in Figure 1. Incoming students in 2020 scored slightly lower on the history test; as discussed earlier, this was due to a decreasing linear trend, not to an unexpected drop in 2020. The 2020 students had considerably larger variability among their scores compared to students from the previous years, except in global issues. Typically, students with AP credit in US history or political

science scored the highest on the history test, followed by those with transfer credit, and then those with no credit. A similar pattern is observed for scientific reasoning. Overall, this pattern holds for 2020. For the global issues assessment, very few students had AP or transfer credits so we did not separate the students into subgroups. As in Figure 1, the largest differences in Table 2 between 2020 and previous years are found on the scientific reasoning assessment. Students in 2020—regardless of whether they had previous credit or not—scored distinctly lower than those in previous years with much higher variability among the scores, particularly on scientific reasoning. Could the interaction between cohort and assessment subject be due to differences in effort? We turn to answering this question next.

Comparable effort was reported by all five cohorts on all assessments, except on the scientific reasoning assessment. Slightly lower levels of effort were reported in 2020 across all assessments; however, they did not seem to deviate much from previous cohorts.

Table 2
Performance across Cohorts

Test	Cohort	Raw Score Mean (*SD*) *N*			
		All	AP	Transfer	No credit
History					
	2020	21.77 (7.26) 841	28.84 (5.81) 57	21.65 (6.50) 80	21.21 (7.16) 704
	2016-19	22.48 (6.39) 4242	30.69 (4.77) 280	22.03 (5.74) 414	21.89 (6.12) 3548
Global Issues					
	2020	16.76 (5.05) 840			
	2016-19	17.40 (4.99) 3767			
Scientific Reasoning					
	2020	38.66 (10.29) 457	45.89 (9.37) 35	38.43 (9.69) 44	38.01 (10.21) 378
	2016-19	44.35 (7.88) 2754	51.65 (6.40) 201	44.54 (7.95) 174	43.72 (7.68) 2379

Note. Subgroup scores are not reported for global issues, because few students had AP or transfer credits in this domain (N = 22 in 2020, N = 64 in 2016-2019). Students were removed if they omitted more than 25% of the items.

Effort

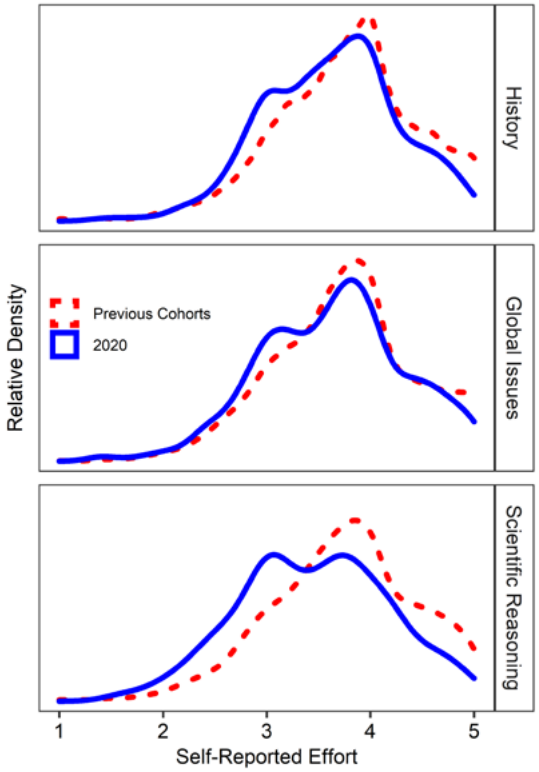
Comparable effort was reported by all five cohorts on all assessments, except on the scientific reasoning assessment. Slightly lower levels of effort were reported in 2020 across all assessments; however, they did not seem to deviate much from previous cohorts (see Table 3). For example, in history, the mean dropped 0.10 (on the 5-point scale) from 2019 to 2020, similar to the 0.13 drop from 2016 to 2017. In global issues, the 2020 mean was only 0.01 below the 2018 mean. Similarly, in scientific reasoning, the 2020 mean was 0.04 below the 2018 mean. These differences are not far from the normal year-to-year fluctuations.

The means and standard deviations, however, do not give a full comparison of effort across cohorts. Refer to Figure 3 for density plots of effort comparing 2020 cohort and previous cohorts combined. The 2020 effort appears bimodal, especially in scientific reasoning. There was a larger bump in students responding *neutral* (3) in 2020. This bump may be related to the greater density of very low scores seen earlier in Figure 2.

Table 3
Means and Standard Deviations of Self-Reported Effort across Cohorts and Assessments

Test	*M* (*SD*) *N*				
	2016	2017	2018	2019	2020
History	3.87 (0.67)	3.74 (0.68)	3.74 (0.67)	3.73 (0.70)	3.63 (0.65)
	1,030	986	1,009	1,164	828
Global Issues	3.69 (0.68)	3.82 (0.67)	3.65 (0.68)	3.73 (0.68)	3.64 (0.70)
	819	1,140	763	1,027	825
Scientific Reasoning	3.86 (0.68)	3.87 (0.70)	3.50 (0.65)	3.71 (0.73)	3.46 (0.72)
	772	667	745	441	456

Figure 3
Density of Self-Reported Effort Scores



To examine the possible relationship between effort and test performance, we have computed the squared correlations between effort and test scores. The squared correlation measures the amount of variation in test scores that can be attributed to exerted effort as reported by the students (see Table 4). Generally, effort seems to be most strongly associated with the scientific reasoning test across cohorts. We also observe an increase in the amount of variation in test scores that is explained by effort in 2020. It appears, overall, that higher levels of effort were associated with higher test performance, especially on the scientific reasoning test.

Time Spent Testing

Another measure of effort is the time students spend taking the test. For each test in 2020, the total time the student spent viewing the test, including short videos at the beginning with information about the test, was recorded. In Figure 4, the standardized score is plotted as a function of the total testing time. Students with transfer or AP credit are not shown.

Table 4
Squared Correlation between Test Score and Self-Reported Effort

Test	Cohort				
	2016	2017	2018	2019	2020
History	.05	.07	.07	.07	.12
Global Issues	.02	.06	.10	.07	.09
Scientific Reasoning	.11	.10	.17	.17	.21

Do students tested remotely in 2020 show less or greater likelihood of correctly answering specific items on the tests than students tested in person, after controlling for ability?

The relationship between time and score appears to be non-linear, especially in science. For students who spent little time on the test, scores increased as time increased. For students who spent at least moderate amounts of time testing, there was little relationship between time and score. Only the first 30 minutes are shown in Figure 4; after that point, the lack of relationship between time and score continued. A regression line was fit to the relationship between the natural log of time and scores. The analysis for fitting the regression line included students not shown in the graph, beyond the 30-minute point. However, students were omitted from the analyses if their time was more than three times the median testing time; it did not seem plausible these students were spending that much time actually focused on the test. This impacted 4.1%, 5.7%, and 3.7% of the students on the history, global issues, and scientific reasoning tests, respectively. The regression accounted for 13% of the variance in history, 10% in global issues, and 27% in scientific reasoning. Testing each pair of correlations at $\alpha = .017$ for a Bonferroni-corrected familywise $\alpha = .05$, the history and global issues correlations were each significantly different from the scientific reasoning correlation, but not significantly different from each other. Time spent on the test was a better predictor of performance for the scientific reasoning test than for the other two tests.

In the history and global issues tests, the time spent per item was also recorded. From this, an adjusted time was calculated. First, a median time was calculated for each item. When a student spent more than three times the median time on an item, the student's time for that item was replaced with an imputed time[3] and the total testing time was recalculated (here labelled the adjusted time). The log of the adjusted time accounted for 21% of the variance in test scores for both history and global issues. The scientific reasoning test might have shown a comparable increase in the correlation, but item-level response times were not available for this adjustment.

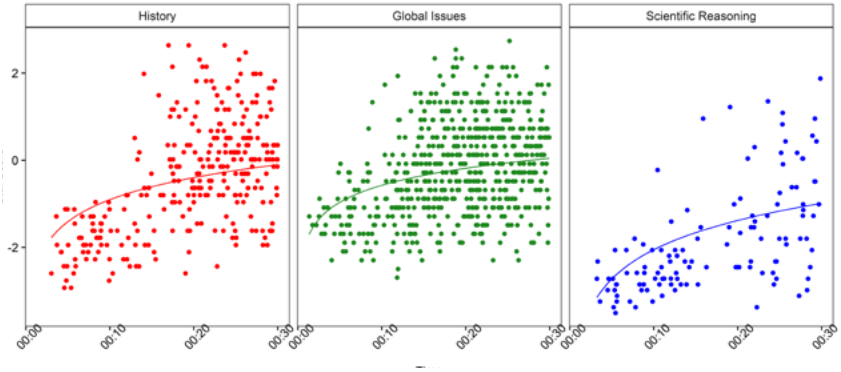
Differential Item Functioning

Remote testing appears to impact students' average performance specifically on the scientific reasoning test. This raises the question of whether remote testing could produce differences not just at the test level, but also at the item level. Do students tested remotely in 2020 show less or greater likelihood of correctly answering specific items on the tests than students tested in person, after controlling for ability? We conducted a differential item functioning (DIF) analysis to examine whether individual items exhibit differential performance between the past four cohorts (2016-2019 combined) and the 2020 cohort after controlling for ability or knowledge level.

We utilized the Mantel-Haenszel method (Holland & Thayer, 1988) to calculate α_{MH}, which is a ratio of the odds of answering an item correctly for the reference group (i.e., past cohorts) over the odds of answering an items correctly for the focal group (i.e., 2020 cohort).

[3] The student's median response time was first estimated as the median across items, excluding any item on which the student took more than 3 times the group's median time for that item. Then the student's ratio was calculated as the ratio of the student median to the group median (overall, across items). Finally, for the excluded items, the response time was imputed as the student's ratio times the group's item-specific median for that item. For example, student Q's median response time across items was 10% more than the group median. On item W, student Q took a break and spent 600 seconds on the item, when the group median time was 22 seconds. Student Q's time for item W would be adjusted to 1.1*22 = 26.4 seconds.

Figure 4
Correlation between Testing Time and Test Score



Note. The fitted line shows the regression of test score on the natural log of time spent. The points cluster closer to the line for scientific reasoning than for the other tests.

The Mantel Haenszel procedure statistically tests the null hypothesis that $\alpha_{MH} = 1$, indicating that the odds for the reference group and focal group are the same. We controlled for false positive rate using the Benjamini-Hochberg procedure (1995). To estimate the effect size of the DIF, we employed ETS classification (Zwick, 2012), which uses the index Δ_{MH} : $\Delta_{MH} = 2.35 \ln(\alpha_{MH})$. According to the ETS classification, an item is classified level A DIF if the absolute value of Δ_{MH} is less than 1 **or** if Δ_{MH} is not statistically significantly different from 0. To be classified as level C DIF, an item has to show an absolute value of Δ_{MH} that is equal to or greater than 1.5 with a Δ_{MH} that is statistically significantly different from 1. Level B classification includes items that do not meet level A or C requirements.

JMU's remote Assessment Day was an exceptional opportunity to study performance differences attributable to testing settings (in-person versus remote) in low-stakes, student learning assessment.

For the history test, item 8 and item 13 showed level C DIF. For the global issues test, none of the items showed DIF with a large effect size (i.e., $\Delta_{MH} \geq 1.5$). For the scientific reasoning test, only item 33 was identified as exhibiting level C DIF. All three items favored the reference group (i.e., previous cohorts) over the focal group (i.e., 2020 cohort). That is, after matching the 2020 examinees with examinees from the previous cohorts with the same total scores, the previous cohorts scored higher on these three items. Inspecting the content of said items, we could not find any plausible explanation as to why these items functioned differently. The lower performance on the scientific reasoning test in 2020 seems to be a pervasive effect, not limited to specific items.

Conclusion

JMU's remote Assessment Day was an exceptional opportunity to study performance differences attributable to testing settings (in-person versus remote) in low-stakes, student learning assessment. The results from the remote Assessment Day were contrasted with results from the previous four cohorts tested in person to control for any pre-existing trends. In terms of mean performance, students tested remotely in 2020 followed the preceding trend of decreasing scores on the history and global issues tests. However, the 2020 cohort exhibited significantly lower scores on the scientific reasoning test than their counterparts in previous years. Those students also showed a different distribution of effort on the scientific reasoning test than students in previous cohorts due to lower effort levels produced by a subgroup of the 2020 students, producing a bimodal distribution. Test performance on scientific reasoning also exhibited this shift in distribution. The scientific reasoning test was longer than the other two tests (66 items vs. 40 and 31), and science may be perceived as more difficult by students. Thus, the different patterns of effort and performance may be attributable to the higher cognitive demand of the scientific reasoning test. Effort was also measured in the 2020 cohort as the time spent taking the test, which predicted performance better for the scientific reasoning test than for the other tests. In future work, as suggested by an anonymous reviewer, we plan to look further at the group of students who gave reasonable effort to assess how their test performance compares to previous cohorts.

We also assessed whether the observed score differences were consistent across items or if instead there was differential item functioning (DIF). Only three items showed large and significant DIF effects between the 2020 cohort and previous cohorts. Evaluating the content of those items judiciously did not yield a reasonable explanation for the DIF.

Overall, the results from JMU's remote Assessment Day suggest that the differences in performance in low-stakes educational assessments observed in students who tested remotely in 2020 can be mainly ascribed to differences in test types. The more arduous scientific reasoning test was the only test showing a significant drop in scores compared to the history and global issues tests which may have required less exertion of cognitive resources. Our findings also highlight the promising potential of remote, large-scale assessment. While a main disadvantage of conducting assessment remotely seems to manifest in the differential performance and effort based on test type, some advantages include less demand for resources (e.g., hiring proctors, reserving rooms, etc.) and the opportunity to collect item-level data on effort. Collecting item-level data allows us to better assess how much effort a student put forth on a test as evidenced by time spent on each individual item rather than the test as a whole. We plan to apply the same remote administration procedures of the 2020 Assessment Day to at least one more assessment day at JMU to further examine the effect of test type and effort levels using data collected at the item level.

We recognize a few limitations of the current study. Effects of remote testing in 2020 may have been impacted by anxiety or other construct-irrelevant factors besides effort due to the pandemic. Lower scores exhibited by the students may have also been affected by events in the semester previous to their enrollment at JMU, when secondary school classes were abruptly moved online. Nonetheless, the current results provide insight into some factors that may impact remote testing. We will continue to study those factors as we assess the same 2020 cohort after completing 45-70 credit hours.

AUTHORS NOTE:
Sarah Alahmadi, https://orcid.org/0000-0002-9985-6807

Christine E. DeMars, https://orcid.org/0000-0003-0050-3655

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* (Series B), 57, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77. https://doi.org/10.1207/s15324818ame1301_3

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In Wainer H & Braun HI (Eds.), Test validity (pp. 129–145). Hillsdale, NJ, US.

Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education*, *7*(1), 271-294. https://doi.org/10.1111/j.1540-4609.2008.00220.x

Huff, K., Cline, M., & Guynes, C. S. (2012). Web-based testing: Exploring the relationship between hardware usability and test performance. *American Journal of Business Education (AJBE)*, *5*(2), 179-186. https://doi.org/10.19030/ajbe.v5i2.6820

Jankowski, N. A. (2020, August). *Assessment during a crisis: Responding to a global pandemic*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/08/2020-COVID-Survey.pdf

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, 29(4), 555-572. https://doi.org/10.1007/s10869-014-9343-z

Pastor, D. A., Foelber, K. J., Jacovidis, J. N., Fulcher, K. H., Sauder, D. C., & Love, P. D. (2019). University-wide assessment days: The James Madison University model. *The Association for Institutional Research (AIR) Professional File*, *144*, 1-13.

Pastor, D., & Love, P. (2020). University-wide assessment during Covid-19: An opportunity for innovation. *Intersection*: *A Journal at the Intersection of Assessment and Learning*, *2*(1), 17617.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, *31*(4), 226-241. https://doi.org/10.1080/08923647.2017.1258628

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14*, 8-9.

Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior*, *24*(3), 1216-1228. https://doi.org/10.1016/j.chb.2007.04.006

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x