Abstract

Universities administer assessments for accountability and program improvement. Student effort is low during assessments due to minimal perceived consequences. The effects of low effort are compounded by assessment context. This project investigates validity concerns caused by minimal effort and exacerbated by contextual factors. Systematic disruptions that affect effort impact the validity of scores. Effort and scores from four administrations of James Madison University's (JMU) remote Assessment Day were examined; these semesters presented unique, changing contexts. Special attention was paid to Spring 2022 which had numerous contextual factors (e.g., online assessment, campus suicides) affecting students and their assessment environments. Time spent testing varied across semesters mirroring the varied scores. With one exception, our results showed lower effort in Spring (posttest) than Fall (pretest) assessments which led to estimates of little or no gain between pretest and posttest. Implications and limitations are discussed.

AUTHORS

Kelsey Nason, M.A.
James Madison University

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

The Impact of External Events on Low-Stakes Assessment: A Cautionary Tale

Universities assess student learning outcomes in general education programming in one of two ways: course-embedded data collection and low-stakes assessment. Course-embedded assessment can require consistent, considerable amounts of work on the part of faculty; this may include training for rating assignments on a common rubric or time harvesting specific assignments from course syllabi. However, students tend to be more motivated to do well as there are more personal consequences like grades (Wise & DeMars, 2006). Low-stakes assessment can be facilitated through a central-body at the university, eliminating a large portion of the persistent work for faculty. However, with no personal consequences, students are less motivated to put forth their best effort (Wise, 2019). This paper investigates the validity concerns that develop due to low effort, a side effect of low-stakes assessment, and additional contextual factors: different environmental conditions that impact test-taking effort and subsequent scores. Any systematic disruptions or factors that affect assessment may impact the validity of scores. Such effects, in turn, change the way we interpret scores and can impact programmatic and/or institutional decisions (Finn, 2015).

CORRESPONDENCE
Email
nasonkt@jmu.edu

Validity indicates whether the interpretations of scores are supported by evidence for the proposed uses of tests (Benson, 1998). The test developers and score users want to interpret the scores as indicators of some intended construct, such as achievement of specified learning outcomes in a content area (e.g., information literacy). Anything outside of the intended construct that influences test performance is labelled construct irrelevant. If construct irrelevant sources systematically affect scores, but the subsequent interpretations

are only in terms of the intended construct, the interpretations will be invalid. The scores might measure the effects of a different construct altogether. For example, suppose students take a math test in a hot room. Scores from this math test may be more indicative of how well students can focus in such temperature conditions rather than their math knowledge. Eliminating the hot temperature condition allows the observed scores to better isolate the construct of interest: math knowledge. If decisions are based on these score interpretations, ones involving contextual elements like the hot classroom used during a math test, it is important to acknowledge when and how contextual factors, or validity concerns, may be impacting understanding of scores.

Construct irrelevant variance can be problematic in low-stakes assessment; these testing conditions are especially vulnerable to validity concerns that arise from context because results are often impacted by student effort and student effort is further influenced by context. Because students tend to put less effort into low-stakes assessment, many low-stakes assessments produce results that are not reflective of true ability or knowledge (Wise, 2019); in fact, they are often underestimations of student ability in a particular subject (Wise & DeMars, 2005). Scores will be attenuated due to low effort exertion. Effort has been reported to attenuate other value-added indices (Finney et al., 2016).

Effort can change the results we get from low-stakes testing and impact our interpretation of gain in scores amongst students across levels and administrations (Rios, 2021; Wise & DeMars, 2010). In Rios et al. (2017), researchers used simulated data to determine if responses lacking effort would underestimate aggregated scores on an assessment. The researchers found this to be the case: respondents with low effort in the simulated study caused attenuated score means. With real data, as opposed to simulated data, there are a variety of ways researchers can measure effort across different testing occasions. One method is through self-report measures. Sessoms and Finney (2015) used the Student Opinion Survey (SOS) to measure effort in college students on low-stakes assessment over time with all other testing characteristics held constant. They found that the average effort declined across test administrations. Another method to measure effort is response time: short amounts of time spent either on the total test or on individual items may be considered indicative of low effort. Using response times as a measure of effort, Yildrum-Erbasloi and Bulut (2020) conducted a study to see how effort can moderate gain estimates using a large-scale, low-stakes reading assessment administered to elementary school students in the Fall and Spring. After filtering slow-responding students and rapid-guessing students, both patterns indicative of low effort, they found that score gain estimates for students significantly increased. They suggested this indicated that score gain estimates of students before filtering non-effortful responses were deflated.

Contextual factors can further impact student effort. They may disrupt student focus, mood, and concentration. A recent example of an external event that had an impact on assessment and higher education at large is the COVID-19 pandemic. Not only were universities expected to transition what were once in-person, proctored assessments to online platforms, but students were also expected to deal with the potential distractions, socio-emotional concerns, and connectivity issues of remote schooling. James Madison University (JMU) was no exception to this changeover. However, in addition to conducting course- and program-level assessments remotely, JMU also had to continue its university-level Assessment Day program that has cultivated over 30 years of longitudinal data on student proficiencies in different learning outcomes.

Contextual factors can further impact student effort. They may disrupt student focus, mood, and concentration.

The specific procedures behind Assessment Day at JMU are documented in Pastor et al. (2019). The event usually involves around 4,000 students at the beginning of the Fall semester (first-year students) and the Spring semester (students who have obtained 45-70 credit hours) to create a pretest posttest design. The assessments administered during Assessment Day are considered low-stakes as students do not face personal consequences based on their performance. With the effects of the pandemic, this event was moved from its typical in-person, proctored, paper-and-pencil design to an online, un-proctored platform with many modifications to its typical procedures (Pastor & Love, 2020).

In this paper, we compare the results from different online assessments administered in different semesters to different cohorts of students over the course of three years.

The Fall 2020 Assessment Day was the first remote assessment day, and students were sent away from campus after the first few days of the semester while testing was still ongoing due to a rise in COVID-19 cases on campus. During Spring 2021 Assessment Day, many classes were still remote, and students were adapting to hybrid class formats. A study was conducted to see the performance-related effects of this switch. Alahmadi and DeMars (2022) reviewed JMU Assessment Day results from five cohorts of incoming students that overlapped the pre-pandemic and online administrations of assessments. They found that remote assessments during the pandemic yielded lower student effort and performance, particularly in one of the more cognitively demanding tests that was administered. It seems likely that much of the decrease in scores was due to the change in context rather than changes in student knowledge. As follows, score interpretations based solely in terms of the intended construct would be of questionable validity.

The pandemic was an environmental factor that had an effect on both practitioners and students; this impact was seen in Alahmadi and DeMars (2022). As COVID-19 procedures slowly dwindle, one might expect test performance to return to earlier trends. Thus, we were hoping that Spring 2022 performance would be higher than Spring 2021. However, there were other external factors—alternative contexts—that impacted students and, subsequently, test scores. An extreme, tragic example of this affected Assessment Day at JMU during the Spring 2022 semester. In addition to students continuing to adjust back to in-person classes in the wake of the pandemic, JMU experienced a great deal of loss. Specifically, just days before Assessment Day, students were faced with news of a fatal campus shooting at a nearby institution followed by more than one suicide on JMU's campus; one of these suicides was witnessed by students. In response to these traumatic events, an announcement was disseminated 'cancelling' Assessment Day. The message was redacted a day later, noting that student participation was still required but the date for completing assessments was extended. This left students very confused about their participation while grieving the loss of fellow community members. Although deadlines were extended and communications with students were increased, this external event was expected to have an impact on the results of Assessment Day.

The purpose of this study was to examine if student effort and assessment performance varied over time, potentially complicating the interpretation of cross-cohort comparisons with increased attention to the cohort that had their posttest in Spring 2022. In this paper, we compare the results from different online assessments administered in different semesters to different cohorts of students over the course of three years. This period presents a unique opportunity to explore how constantly changing context may impact large-scale, low-stakes assessment. We investigated the following research questions:

1. Did students in different semesters differ in how long they spent on the tests?
2. Did students in Spring 2021 and students in Spring 2022 differ in their test scores? Did students in different Fall semesters differ in their test scores?
3. For students who took the same test in Fall 2020 and Spring 2022, are differences in time spent testing related to score gains from pretest to posttest?

Time spent on the assessments is viewed as a measure of effort. We expected time variation in these semesters due to students enduring different contextual factors; in addition, we expected more students in Spring semesters to exert low effort due to their second-year status. These students have historically tried less on these low-stakes assessments (Sessoms & Finney, 2015). Making comparisons between different Spring semesters, and separately between different Fall semesters, allows us to separate other contextual effects from the confounding context of Fall vs. Spring. For many students, COVID-19 had a smaller impact on life in Fall 2022 than in Fall 2020, so there may have been a smaller proportion of students exhibiting non-effortful testing times in the Fall 2022 cohort than the Fall 2020 cohort. Context impacted students differently in Spring 2021 and Spring 2022; this might have led to more or fewer students with non-effortful times.

In addition to differences in effort, we expected scores to vary depending on contextual factors as well. More specifically, we expected there to be differences between the Fall semesters showing higher scores in Fall 2022 compared to Fall 2020 which was disrupted by COVID-19. In Spring 2022, we initially expected to see higher scores than Spring 2021 due to recovery from the COVID-19 disruption. However, when unanticipated extreme circumstances surrounded the Spring 2022 administration, our expectations changed.

Like the findings of Rios et al. (2017) and Yildrum-Erbasloi and Bulut (2020), we expected that score gains would be affected by the time spent testing. Students who expend little effort on the posttest could be deflating gain estimates between pretest and posttest assessment results. Conversely, if any students expended effort on the posttest but not the pretest, gain estimates could be inflated.

Method

Participants

Participants were first-year and second-year[1] students entering or continuing their time in the university between 2020 and 2022. All of these students were required to participate in Assessment Day, but they were randomly assigned to take different sets of assessments to complete their assessment requirement. These students follow the university's general demographic statistics which report a female to male ratio of 59:41% and roughly 78% of students identify as white (James Madison University, 2022). This study used data collected from students who completed at least one of the three tests described below. For each test, there were anywhere from 500 to 1,000 student scores used in analysis. Students were given a two-week period to test in the Fall and one day in the Spring; students received the links to their assessments in an email and could complete them in their chosen environment (dorm room, library, computer lab, etc.). There were no consequences for not participating in Fall 2020, Spring 2021, and Spring 2022. In the other semesters, a registration hold was placed on student records if they missed the assessment deadline; after they completed their assigned assessments, the hold was removed.

Assessment Instruments

Three assessments administered to assess General Education knowledge were used in this study, because they were administered for at least two semesters between Fall 2020 and Fall 2022. Each assessment is of a different length and different subject. The assessments were developed by university faculty to target knowledge in history (40-item measure), global processes (31-item measure), and information literacy (30-item measure). These tests were consciously created with no essay questions or other more cognitively-taxing question formats; these types of dynamic questions, in contrast with more simple question formats like multiple-choice, require more effort from students (DeMars, 2000). Items contained four-to-five answer choices. Assessments are randomly assigned to students in different sets. Each set contains three to four assessments that are a mix of cognitive and non-cognitive tests. The assessments of interest in this study were all cognitive. Test sets are not consistent across semesters—that is, assessments analyzed in this study were not administered in the same order nor mixed with the same cognitive and non-cognitive assessments each semester. This was potentially another contextual factor that impacted score validity.

Students who expend little effort on the posttest could be deflating gain estimates between pretest and posttest assessment results. Conversely, if any students expended effort on the posttest but not the pretest, gain estimates could be inflated.

Four semesters, two Fall sessions and two Spring sessions, were used in analyses for the U.S. history assessment and the global processes assessment. Two semesters, a pretest and posttest for a single cohort, were used in analyses for the information literacy assessment because it was not administered in the other two semesters.

[1] We use the descriptor second-year for brevity: this group includes students with 45-70 credits before the Spring semester. The group thus includes some 1st year students who took college credit concurrently with high school, as well as some 3rd year students who did not earn quite enough credits to be tested in their 2nd year.

Time Spent Testing

Of note is the abnormal behavior exhibited by students in Spring 2022. This semester shows the largest amount of non-effortful responses creating a nearly bimodal distribution.

With the continued online format of Assessment Day, JMU can collect time information on its assessments. As suggested by Wise and DeMars (2005) and Yildrum-Erbasloi and Bulut (2020), these response times—the difference in seconds between when an item is answered and when it was initially presented to the student—can be used as a proxy for effort (Wise & Kong, 2005). For each assessment, the time spent on each item was recorded. Testing time was defined as the sum of the item times.[2] An additional index for measuring effort is response time effort (RTE): the proportion of items on which the examinee's time exceeded some minimal threshold (Wise & Kong, 2005). Common thresholds are 10%, 20%, or 30% of the mean time spent on a given item (Wise & Kuhfeld, 2020). We used 20% of the median time spent on item *i* as that item's threshold.

Results

Time Spent Testing

We used the total time spent testing (in minutes) as a proxy for test effort exerted by students. Extremely short times are indicative of low effort, although once students are within reasonable ranges of testing time, effort is likely unrelated to time. Graphical analyses were used to determine the differences in effort exerted by students on each test during different semesters. Figure 1 shows the density (proportion of students) graphs for each test against time spent testing. The vertical line demarcates students who completed more than 5 items per minute (6 minutes for 30 items, 8 minutes for 40 items). This point is somewhat arbitrary; these students are clearly non-effortful respondents, but students who took just a little more time may not have applied full effort throughout the test.

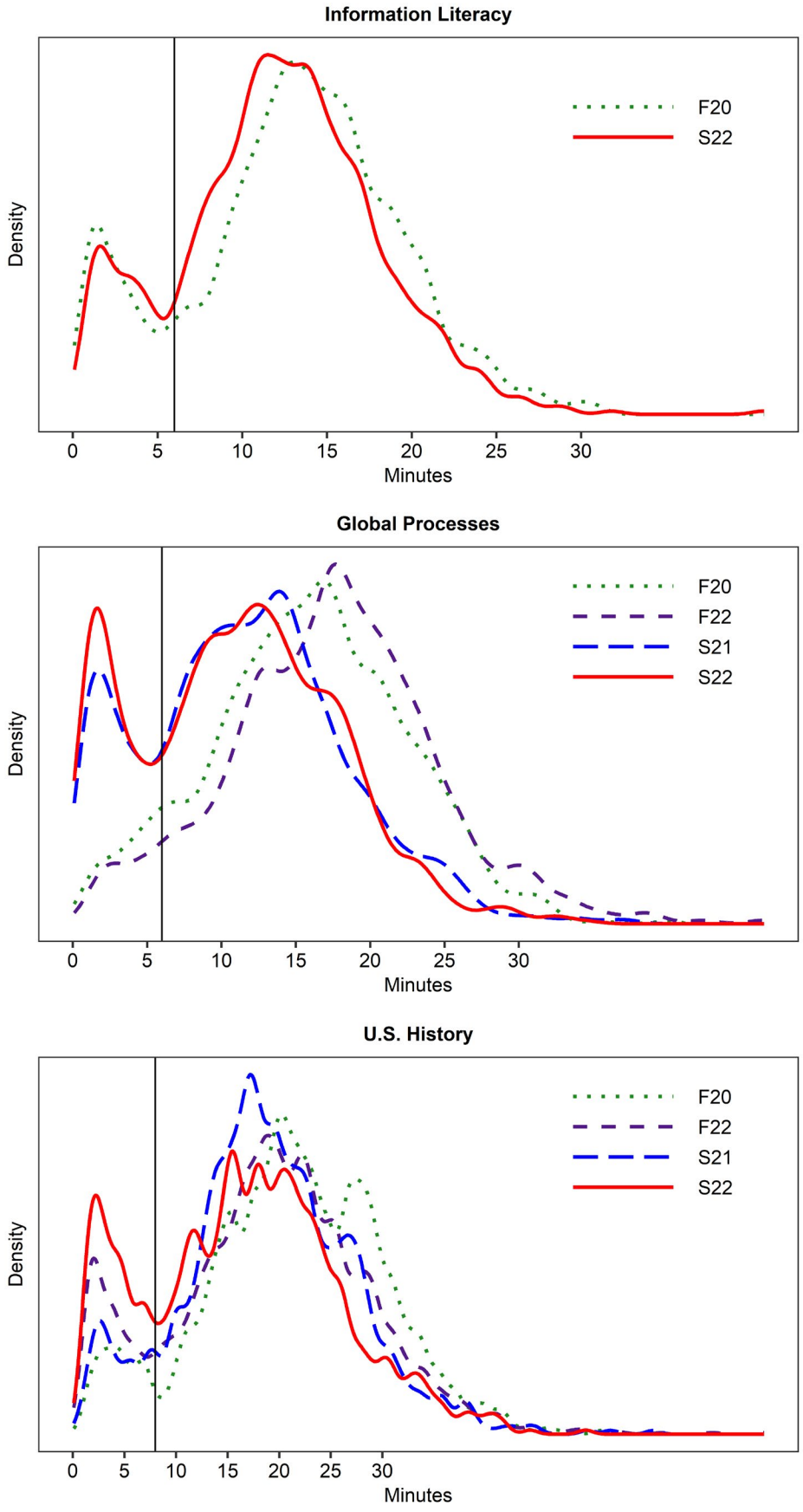
On the information literacy assessment, we looked at time spent testing for students across two semesters: Fall 2020 and Spring 2022. Here, we see that the two semesters have similar proportions of students producing non-effortful responses. Students in Fall 2020 showed a slightly higher proportion of students exerting low effort than students in Spring of 2022. There were no other anomalies of note between these two semesters.

The global processes test told a different story. Most of the four semesters pictured (Fall 2020, Fall 2022, Spring 2021, Spring 2022) show a majority of students spending between 15- and 20-minutes testing which is a reasonable, or effortful, amount of time. We also see that both Fall semesters show smaller proportions of students exhibiting non-effortful behavior compared to the Spring semester students. Of note is the abnormal behavior exhibited by students in Spring 2022. This semester shows the largest amount of non-effortful responses creating a nearly bimodal distribution. We see similar results on the U.S. history assessment; Spring 2022 students show the largest amount of non-effortful responses compared to the other semesters. Interestingly, the next group with the most non-effortful responses were students in Fall 2022 rather than the other Spring semester.

[2] If a student spent an excessive amount of time, more than 120 seconds, on one item, the item response time was adjusted before summing. To make this adjustment, the median time, across students, was calculated for each item. Then for each student *j* and item *i*, the ratio of the response time to the median response time was computed. Within each student, the median of these ratios, across items, was computed after exempting the items with excessively long times. Then for any item *i* with an excessively long response from student *j*, student *j*'s median ratio was multiplied by item *i*'s median response time. For example, if student Q spent 10 minutes on item 3, student Q's median ratio was 1.1, and the median response time for item 3 was 20 seconds, student Q's response time was modified to 22 seconds before computing total testing time. This adjusted time consistently had higher correlations with test scores than unadjusted total time, presumably because the student was not focused on the item for the entire time recorded.

Figure 1

Density graphs of total time spent testing on the Information Literacy, Global Processes, and U.S. History assessments across different semesters.



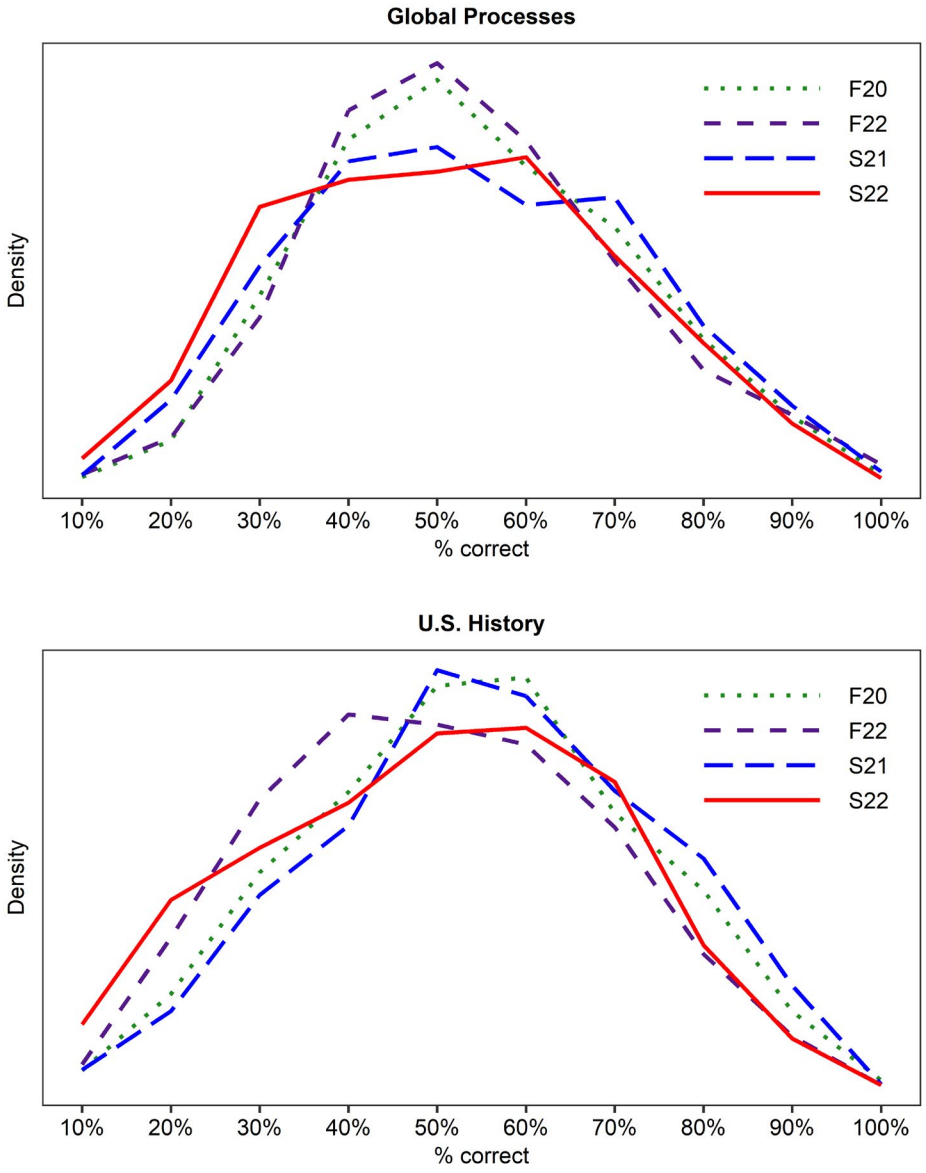
Test Scores

Like the dip in effort we observed in time spent testing, students performed worse in Fall 2022 than the previous Fall 2020

A series of analysis of variance (ANOVA) tests looked at the differences in percent correct on the assessments in addition to graphical analysis. Figure 2 displays percent correct to track changes from semester to semester. The focus here is on comparing different cohorts of students at the same point in their academic careers (Fall 2020 vs. Fall 2022 or Spring 2021 vs. Spring 2022); differences over time will be addressed later.

The global processes scores, like time spent testing, told a unique story. There was not a significant difference between scores of Fall 2020 (M= 0.54, SD=0.16) and Fall 2022 (M= 0.52, SD=0.16), $F(1, 1578) = 2.23$, $p = .136$. The same analysis was run to compare Spring 2021 (M= 0.53, SD=0.18) and Spring 2022 percent correct (M= 0.50, SD=0.19); a significant difference was found showing Spring 2021 yielded higher scores than did Spring 2022, $F(1, 1707) = 10.52$, $p = .001$. Looking at the graph, we see that students in Spring 2022 showed the highest number of students obtaining low test scores. Their mean score was worse than students who just entered the university in the Fall semesters. In addition, in Spring 2022 a lower proportion of students obtained high test scores; again, their performance dipped below that of students in the Fall semesters.

Figure 2
Density graphs of test scores for Processes and U.S. History Assessments across four semesters.



We observed similar patterns of behavior in students who took the U.S. history assessment with some notable differences. First, there was a significant difference between Fall 2022 semester scores (M= 0.49, SD=0.13) and Fall 2020 semester scores (M= 0.54, SD=0.18), $F(1, 3006) = 45.40$, $p < .001$. Like the dip in effort we observed in time spent testing, students performed worse in Fall 2022 than the previous Fall 2020; we see this in the graph as Fall 2022 semester scores peaked earlier with a more rounded distribution than the distribution of Fall 2020. The same analysis was run to compare the Spring semesters; a significant difference was found with Spring 2021 (M= 0.56, SD=0.18) scores higher than Spring 2022 scores (M= 0.51, SD=0.19), $F(1, 1707) = 10.52$, $p = .001$. Like global processes scores in Spring 2022, many students scored low compared to Spring 2021 and compared to the Fall semesters. There were also fewer high scores obtained by students in Spring 2022 than all other semesters in the graph. Gain Scores

The relationship between the gain in scores (difference between the Spring 2022 scores and Fall 2020 scores) and the difference in response time effort[3] (RTE) between Spring 2022 and Fall 2020 for students in each test was looked at graphically (see Figure 3). First, ANOVAs were run to look at the differences in scores between the pre- and posttest for each assessment administered to this cohort. A significant difference in scores was found in information literacy scores with Spring 2022 having a higher percent correct than Fall 2020 students, $F(1, 1803) = 21.91$, $p < .001$. At face value, these results reflect student learning from exposure to general education programming. For the global experience test, a significant difference was found between scores in Fall 2020 and Spring 2022 showing that the Fall scores were higher than the Spring scores, $F(1,1801)=16.63$, $p<.001$. In a similar manner, a significant difference was found between scores in Fall 2020 and Spring 2022 on the U.S. history test showing students performed better in the Fall than the Spring $F(1,1772)=17.97$, $p <.001$. Between the two administrations, it appears students lost knowledge. This is different from previous years, in which students showed an average gain as they progressed through the university.

To further examine these differences, gain scores were plotted on the y-axis and the differences in RTE were plotted on the x-axis. If students scored better on the posttest, all points would be above the origin on the y-axis; higher scores in the Spring are evidence of student learning. If students put in equal effort during the Fall and Spring semesters points would be clustered around the origin of the x-axis. Any deviation from this area becomes a validity issue as effort can start to affect subsequent interpretations.

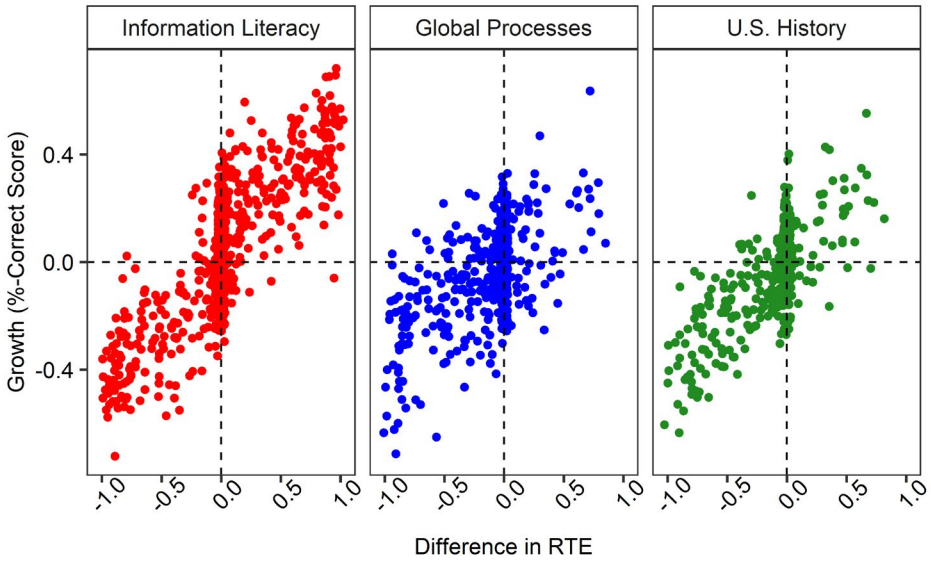
A significant difference in scores was found in information literacy scores with Spring 2022 having a higher percent correct than Fall 2020 students… At face value, these results reflect student learning from exposure to general education programming.

For the information literacy assessment, we found a significant correlation between gain scores and pre-post differences in RTE (r=.82, $p < .001$). This was also the case for the global processes assessment (r=.59, $p < .001$) and the U.S. history assessment (r=.73, $p < .001$). This significant relationship indicates that generally, students who exerted lower effort on the posttest than the pretest had lower (generally negative) gains from pre to posttest. Across all three tests, students in the top right or lower left quadrants represent a validity threat. In the lower left quadrant, students gave more extremely rapid responses on the posttest; in the top right, students gave more extremely rapid responses on the pretest. In the lower left quadrant, it appears that many of the students lost knowledge between the first (Fall) and second (Spring) administrations of the assessments. In the upper right quadrant, the students appear to have unrealistic gains. Note there are more students in the upper right quadrant than the lower left which likely explains the average decrease in scores over time.

[3] We selected response time effort (RTE) instead of total time spent testing because, among students who give effortful responses, total time may be lower in Spring due to higher levels of knowledge. Thus, slightly shorter testing times would not indicate lower effort. RTE, in contrast, measures the proportion of items to which the student gave an effortful response. A response is labelled effortful if the student spent at least 20% of the median time on the item. RTE is calculated by dividing the number of items during which a student exerted effort by the total number of items.

Figure 3
Graph depicting differences in gain in scores and RTE between Fall 2020 and Spring 2022.

In the lower left quadrant, it appears that many of the students lost knowledge between the first (Fall) and second (Spring) administrations of the assessments. In the upper right quadrant, the students appear to have unrealistic gains.



Note: The y-axis shows the difference in gain in scores while the difference in response time effort is on the x-axis for each assessment (information literacy assessment, global processes assessment, and U.S. history assessment). Because many points were layered on one another, they were jittered slightly.

Discussion

JMU's remote Assessment Day provided an opportunity to study how contextual factors, coupled with low effort, a common feature of low-stakes assessment, created validity concerns in assessment and score interpretation. Specifically, we looked at three different assessments administered for at least two semesters over the past three years to examine effort exerted (proxied by time spent testing), percent of correctly answered questions, and the relationship between gain in scores and RTE from the pretest Fall administrations to the posttest Spring administrations. During these Assessment Day administrations, there were numerous events that impacted students and subsequently their testing experiences (e.g., COVID-19 pandemic in Fall 2020 and Spring 2021, loss in Spring 2022). We expected these events to impact scores and, as a result, the validity of these scores.

Two additional factors were considered when interpreting our results: the students' year in university and the order of tests. Students further along in the academic program (in this case, students with 45-70 credit hours) generally report lower effort on low-stakes assessments (Eklöf et al., 2014; Sessoms & Finney, 2015; Thelk et al., 2009) and show more rapid-guessing (Wise & DeMars, 2010). Zilderberg (2013) suggests this is due to more discontent among students further along in the program. For conciseness, we will label this the *Sophomore effect*. The order of assessments may also make a difference in effort. Students may be more cooperative on the first test and spend more time and effort on it. In subsequent tests, students may exhibit boredom or lack of interest attenuating their effort on test items (DeMars, 2007; Deribo, Goldhammer, & Kroehne, 2023). Previous research has suggested this phenomenon within an assessment rather than across assessments (e.g., Wise, 2006; Wise, Pastor, & Kong, 2009).

We found that in the global processes and U.S. history assessments, Spring 2022 students exerted lower effort than all other semesters (notably, lower than when these students took the assessment as incoming first-year students). As a reminder, Spring 2022 housed multiple traumatic, contextual factors (e.g., campus suicides, nearby shooting). We

largely attributed lower effort to the circumstances surrounding this administration. However, for information literacy, we did not find a drastic number of students exerting low effort during Spring 2022 compared to Fall 2020 students. The information literacy test was given second (after global processes) in the Fall but first in the Spring. Thus, the Sophomore effect may have been mitigated by the opposite effect of test order. The global processes, in contrast, was administered first during both Fall semesters but second or third in the Spring semesters. The effects of test order and the Sophomore effect may have compounded to yield particularly large differences between Fall and Spring effort. The U.S. history test was administered first during each semester except Fall 2022 where it was second to either the global processes test or a more taxing environmental reasoning test. The effects of order are likely why we see low effort in Fall 2022 while the Sophomore Effect and contextual factors compounded to produce the low effort seen in Spring 2022.

A similar pattern emerged in scores as well. We saw some difference in scores Fall to Fall in the U.S. history assessment, but not in the global processes assessment. The difference between Fall scores is likely due to the order of administration of the U.S. History test in Fall 2022. We also saw significant differences in scores from Spring to Spring for both tests. Specifically, we saw more students in Spring 2022 showing extremely low scores and fewer students with high scores than any other semester. As a result, it looked as though students knew less in the Spring semester than they did 1.5 years before. This is evidenced by our look into the relationship between score gain and change in RTE which yielded significant, positive correlations. Mainly, this showed us that applying effort on tests translates to more gain in scores; unfortunately, we saw a lot of students not exerting equal effort in both semesters and their scores seemed to decrease between Fall and Spring administrations. As these students continued to be successful at the university, it is doubtful they lost knowledge like these scores suggested. This variance in changes in effort illustrates that systematic changes in the testing context, such as local or global events, testing order, and progression through coursework, do not influence the effort of all students equally. Although most students show either no change or less effort in their second year than their first year, some students show the opposite pattern. On average students become fatigued or less cooperative on tests administered later in the sequence, but the effect is not uniform. Similarly, the effort of some students is impacted more than others by external events.

There are limitations to this study of time, scores, and gain scores with RTE in the wake of contextual factors. First, we assume time spent testing is a good proxy for effort exerted on assessments. Although the literature supports this assumption, it is not an exact measure of effort but simply a way of flagging students who exerted almost no effort. Sometimes researchers will also employ a self-report measure to use as an additional support for effort exertion during low-stakes assessments (Wolf & Smith, 1995). In addition to the measure of effort, we are unable to specifically identify the environmental element which accounts for variance in effort. A strong argument can be made for the contextual factors, like the circumstances surrounding the Spring 2022 administration and test order, however, this cannot be exactly parsed out. We are unable to definitively state one factor impacts students more than the other. Many other contextual factors could have been present in students' lives accounting for the lapse in effort on assessments.

Our circumstances, although more extreme than typical circumstances, show that external events and other contextual factors can have major consequences on scores and the validity of interpretation.

Validity of scores, or interpretation of scores, is especially important in low-stakes assessment. Our circumstances, although more extreme than typical circumstances, show that external events and other contextual factors can have major consequences on scores and the validity of interpretation. There are always contextual factors that impact students and their assessment environment; some are more personal while others affect larger groups. Practitioners should keep these factors and events in mind when interpreting scores from assessments as these scores can often be attenuated. Test scores will never be an exact reflection of student knowledge. If using an online format for assessments, collecting timing information and looking at time spent testing as a proxy for effort is an easy way to keep effort and validity of scores in mind. If assessments are not online, one could instead use a self-report measure to gauge effort exertion. We hope that sharing the results of our remote Assessment Day through different impactful, external events can provide some information about these unexplored conditions and the importance of context in assessment. Each

institution of higher education will have unique circumstances. Although no other institution will likely share exactly the factors encountered here, practitioners at other institutions can take away the message that a variety of unexpected conditions can have a sizeable impact on effort and test-taking performance which should be considered when drawing inferences about student learning.

References

Alahmadi, S., & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote assessment day. *Research & Practice in Assessment*, *17*(1), 5-15.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational measurement: Issues and practice*, *17*(1), 10-17. https://doi.org/10.1111/j.1745-3992.1998.tb00616.x

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23-45. https://doi.org/10.1080/10627190709336946

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

Deribo, T., Goldhammer, F., and Kroehne, U. (2023). Changes in the speed–ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*, *83*(3). 473–494. https://doi.org/ 10.1177/0013164 4221109490

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, *27*, 31–45. https://doi.org/10.1080/08957347.2013.853070

Facts and Figures. James Madison University. (2022, August). Retrieved March 29, 2023, from https://www.jmu.edu/about/fact-and-figures.shtml

Finn, B. (2015). Measuring motivation in low-stakes assessments. *Educational Testing Service*, ETS RR-15-19. https://doi.org/10.1002/ets2.12067

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

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. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104. doi:10.1080/15305058.2016.1231193

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing*, *15*(4), 356-388. https://doi.org/10.1080/15305058.2015.1034866

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*, *58*, 129–151. https://doi.org/10.2307/27798135

Wise, S.L. (2019). Controlling construct-irrelevant factors through computer-based testing: Disengagement, anxiety, & cheating, *Education Inquiry*, *10*(1), 21-33. doi: 10.1080/20004508.2018.1490127

Wise, S.L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27-41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

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

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. R. (2020). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130-149. https://doi.org/10.1111/jedm.12275

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185-205. https://doi.org/10.1080/08957340902754650

Wolf L. F., Smith J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242. https://doi.org/10.1207/s15324818ame0803_3

Yildirim-Erbasli, S.N., Bulut, O. (2020) The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation*, *26*(7-8), 368-386. doi: 10.1080/13803611.2021.1977152

Zilberberg, A. (2013). *Students' attitudes toward institutional accountability testing in higher education: Implications for the validity of test scores* (Doctoral dissertation). Retrieved from http://www.lib.jmu.edu