Chemistry
Education Research
and Practice

Students who prefer face-to-face tests outperform their online peers in organic chemistry

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-11-2021-000324.R2
Article Type:	Paper
Date Submitted by the Author:	04-Feb-2022
Complete List of Authors:	Beatty, Abby; Auburn University, Biological Sciences Esco, Abby; Auburn University, Biological Sciences Curtiss, Ashley; Auburn University, Chemistry and Biochemistry Ballen, Cissy; Auburn University, Biological Sciences

Students who prefer face to face tests outperform their online peers in organic chemistry

Abby E. Beatty*, Abby Esco, Ash Curtiss, & Cissy J. Ballen
*Corresponding author: aeb0084@auburn.edu
Auburn University, Auburn, AL

In consideration as a Research Article in: *Chemistry Education Research & Practice*

Keywords: computer-based exams; paper-based exams; testing mode; testing mode effect; exams; intrinsic goal orientation; engagement; task value; distance education; academic performance

Conflicts of Interest: There are no conflicts of interest to declare.

Data Availability: All data and code are publicly available at https://github.com/aeb0084/Testing-Modality-in-Organic-Chemistry.git. Code and data are also available here as supplemental files.

Abstract

To test the hypothesis that students who complete remote online tests experience an 'online grade penalty,' we compared performance outcomes of second-year students who elected to complete exams online to those who completed face-to-face, paper-based tests in an organic chemistry course. We pursued the following research questions: (RQ1) Are there performance gaps between students who elect to take online tests and those who take face-to-face tests? (RQ2) Do these two groups differ with respect to other affective or incoming performance attributes? How do these attributes relate to performance overall? (RQ3) How does performance differ between students who reported equal in-class engagement but selected different testing modes? (RQ4) Why do students prefer one testing mode over the other? We found that students who elected to take online tests consistently underperformed relative to those who took face-to-face tests. While we observed no difference between the two student groups with respect to their intrinsic goal orientation and incoming academic preparation, students who preferred face-to-face tests perceived chemistry as more valuable than students who preferred to complete exams online. We observed a positive correlation between performance outcomes and all affective factors. Among students who reported similar levels of in-class engagement, online testers underperformed relative to face-to-face testers. Open-ended responses revealed online testers were avoiding exposure to illness/COVID-19 and preferred the convenience of staying at home; the most common responses from face-to-face testers included the ability to perform and focus better in the classroom, and increased comfort or decreased stress they perceived while taking exams.

Introduction

Negative experiences and performance outcomes in large foundational STEM courses, such as organic chemistry, are frequently cited reasons students leave STEM (Barr *et al.*, 2008; Ost, 2010; Rask, 2010; Seymour and Hunter, 2019). Those who receive low grades are more likely to drop out, and less likely to pursue a STEM degree or enter a STEM field (Mervis, 2011). Thus, research that addresses factors that drive observed performance gaps in organic chemistry has the potential to enhance the persistence and retention of students.

One factor contributing to underperformance in chemistry may be choice of testing mode. Specifically, students who elect to take their assessments remotely online rather than face-to-face and on paper may experience a testing 'penalty'. For this study, testing mode refers to the method of delivering a test to students: either remote online tests (hereafter 'online') or face-to-face paper-based tests (hereafter 'face-to-face'). The testing mode effect refers to differences in student performance between tests given in different testing modes. In our study, students experienced the exact same format of test questions and wrote answers on the same hard copy answer sheets in online and face-to-face testing environments. While we do not seek to untangle the potential effects of *where* students completed their exams, the option to take exams at home (rather than face-to-face) is becoming increasingly common, as online courses surge in popularity. According to the National Center for Education Statistics, in 2018, over one-third of all undergraduate students engaged in distance education, and 13 percent of total undergraduate enrollment exclusively took distance education courses. Of the 2.2 million undergraduate students who exclusively took distance education courses, 1.5 million enrolled in institutions located in the same state in which they lived (Hussar *et al.,* 2020). These values are expected to increase as online learning opportunities are cost-effective and students who are entering higher educations may have families, be involved in part-time or full-time jobs, or have other responsibilities. Online exams are also an integral part of our national efforts to promote diversity, equity, and inclusion, as students who request testing accommodations complete them remotely and the exams are often computer-based.

Some previous work defined testing mode slightly differently, as computer-based or paper-based exams taken in the same environment. One study found that if two students with equivalent competencies completed an assessment in the same testing location, the student who took the paper-based test outperformed the student completing the computer-based test. Specifically, (Backes and Cowan, 2019) examined test scores for hundreds of thousands of K-12 students in Massachusetts and demonstrated a testing mode effect; specifically, they found an online test 'penalty' of approximately 0.10 standard deviations in math and 0.25 standard deviations in English. However, other research disputes these results, with mixed outcomes presented in the literature. Meta-analyses of testing mode effects on K-12 mathematics test scores (Wang *et al.*, 2007) and K-12 reading assessment scores (Wang *et al.*, 2008) demonstrated no statistically significant effect of testing mode. While results on testing mode effects are mixed, less work has been conducted to explain these potential differences. As one of few studies performed to answer this question in undergraduate chemistry settings, Prisacari and Danielson (2017) administered practice tests in the form of computer-based or paper-based assessments to 221 students enrolled in general chemistry and found no evidence of testing mode effects between the two groups. Notably, students were assigned a testing mode based on scheduling availability, not testing mode preference, and

researchers administered all assessments in the same classroom. Researchers concluded that instructors need not be "concerned about testing mode (computer versus paper) when designing and administering chemistry tests."

When given the option of testing mode, some students may simply prefer the convenience of taking college-level exams online from their home. When fifth year medical students were given the *opportunity* to select a testing mode on an exam, researchers evaluated performance differences, the reason behind the choice of the format, and satisfaction with the choice (Hochlehnert *et al.*, 2011). This study did not observe differences in performance based on testing mode. We hypothesize this may be due to the academic maturity of fifth year medical students who participated in the study, but could alternatively relate to the nature of the exam content. Additionally, students who elected to take online exams described their exams as clearer and more understandable.

In this paper, we use the online option in an organic chemistry course to investigate whether differences in grades are reflective of real differences in student performance or of other extrinsic and intrinsic factors that relate to preference for a testing mode. Specifically, we explore how testing mode preference might be related to constructs associated with motivation and engagement, which have been shown to relate to student performance in chemistry (Garcia, 1993; Black and Deci, 2000; Ferrell *et al.*, 2016). We measured two motivation processes, intrinsic goal orientation and perceived value of chemistry (Pintrich *et al.*, 1993). *Intrinsic goal orientation* is motivation that stems primarily from internal reasons (e.g., curiosity, wanting a challenge, or to master the content) (Pintrich *et al.*, 1993). *Task value*, or perceived value of chemistry, is motivation to engage in academic activities because of the students' beliefs about the utility, interest in, and importance of the disciplinary content (Pintrich, 1999). We selected these two distinct constructs because they reflect both intrinsic motivators (i.e., intrinsic goal orientation), such as the desire to develop deeper understanding, and extrinsic motivators (i.e., task value), such as beliefs that the subject material might be relevant to their future careers. If we find that, for example, students who prefer to take online tests have higher intrinsic goal orientation, then one strategy that might work well with online students is to embrace intrinsic factors that motivate them by, for example, encouraging instructors to teach through discovery or problem-based learning approaches. However, if we find that these students display higher level of task value, reflecting those extrinsic factors motivate the students at a higher level, then we may encourage instructors who teach online students to intentionally contextualize course material in real-world examples (e.g., Fahlman *et al.*, 2015).

Another possible explanation for differences in performance between testing modes is their relation to student engagement, an essential part of the learning process (Coates, 2005; Chi and Wylie, 2014). Attempts to measure engagement in undergraduate STEM classrooms come in many forms, including participation and student behavior in the classroom (Sawada *et al.*, 2002; Smith *et al.*, 2013; Chi and Wylie, 2014; Eddy *et al.*, 2015; Lane and Harris, 2015; Wiggins *et al.*, 2017; McNeal *et al.*, 2020; Pritchard, 2008), students' reflections of their own cognitive and emotional engagement (Chi and Wylie, 2014; Wiggins *et al.*, 2017; Pritchard, 2008), and even real-time measurements through the use of skin biosensors (McNeal *et al.*, 2020). Similar to other measures in the current study, we quantified engagement because of its potential power in explaining performance disparities, and

because of past research in the context of undergraduate STEM displaying its importance in academic success and performance (McNeal *et al.*, 2020; Miltiadous *et al.*, 2020).

We expected one of three outcomes: in one scenario, we do not observe testing mode effects. If we do observe a difference in performance, another scenario is that average exam scores are lower among students who elect a particular testing mode primarily due to self-selection effects, where less academically prepared and engaged students tend to prefer one testing mode. Alternatively, students that are equally prepared and engaged may perform at a lower level due to external factors related to a testing mode. To our knowledge, this is the first exploratory study to identify student preferences for testing modes in an undergraduate chemistry setting, and propose explanations for potential differences in performance due to testing modality. We analyzed data from two semesters of an organic chemistry class at a large southeastern university and addressed the following questions: (RQ1) Are there performance gaps between students who elect to take online tests and those who take face-to-face tests? (RQ2) Do these two groups differ with respect to other attributes, such as (a) intrinsic goal orientation, (b) perceived value of chemistry, or (c) incoming academic preparation? How do these attributes relate to performance overall? (RQ3) How does performance differ between students who report equal in-class engagement but selected different testing modes? (RQ4) Why do students prefer one testing mode over the other?

Experimental

Data collection

To address our first research question, we analyzed performance outcomes of students who enrolled in organic chemistry across summer 2020 and fall 2020 semesters at a large, southeastern university (N = 305; **Table 1**). Organic chemistry is an in-depth study of structure, nomenclature, reactions, reaction mechanism, stereochemistry, synthesis, and spectroscopic structural determination. This course was designed for pre-health professionals, science majors, and chemical engineers. Most students take organic chemistry during their second year after completing general chemistry.

The classes in this study met via Zoom three times weekly. Those who took the class in summer met for 75-minute class sessions over 10 weeks, and those who took the class in the fall met for 50-minute class sessions over 15 weeks. Learning Assistants assisted these classes, resulting in a 15:1 student to Learning Assistant ratio. The course design was a flipped classroom model, and in every class session students were randomly assigned to breakout rooms. The instructor uploaded pre-recorded lectures to a Learning Management System and students attended Zoom classes to work through Process Oriented Guided Inquiry Learning (POGIL) handouts and ask questions. For example, a typical class period would begin with 10-15 orientation slides/drawings/pictures followed by 5-10 minutes of general questions. Afterward, students went into breakout rooms to work on questions from the POGIL handout.

Exams typically covered 3-4 chapters of content and were designed to take about 10 minutes per page. Various question types included multiple choice, fill in the blank, or free response. At least 50% of all exam items tested students' ability to use, or interpret, bond line drawings to convey details of organic chemical structures or changes in chemical structure due to reactions. These molecular representations are a commonly used style to simplify complex molecules. Most of these questions are examples base on analogous reasoning (e.g., A+B → C, where either A, B, or C was

missing). Each exam included an opportunity for extra credit, increasing the highest potential score to 120%. Students decided whether they took the online exam or face-to-face exam. The exams were identical in content and distributed at the same time (synchronously). Students who took online exams were proctored by the instructor (ABC) and graduate teaching assistants via Zoom (i.e., no 3rd party proctoring service was used). Students who met face-to-face sat 6-feet apart (socially distanced) in a classroom.

Table 1: Demographic breakdown of students by testing mode (Online or Face-to-Face).

Participants	N=305	
	Online	**Face to Face**
Exam 1	n=143	n=126
Exam 2	n=143	n=127
Exam 3	n=161	n=108
Binary gender		
Women	70.9%	65.7%
Men	29.1%	34.3%
Class Standing		
First year	86.1%	79.8%
Second year	5.6%	11.4%
Third year	5.8%	5.3%
Fourth year	1.8%	1.9%
Post-baccalaureate	0.7%	0.8%
Graduate Student	0.0%	0.8%
Race/Ethnicity		
Asian/Asian American	3.8%	3.6%
Black/African American	3.4%	0.8%
Latino/Hispanic American	4.7%	4.2%
Native Hawaiian or Other Pacific Islander	0.0%	0.8%
White/ European American	75.5%	86.7%
Other	12.6%	3.9%
First-generation student?		
No	73.4%	81.4%
Yes	14.1%	14.1%
Unsure	12.5%	4.4%

The course used Gradescope to evaluate the exams. Students in both formats received equivalent question sets and answer keys. Both testing modes received an "Answer Sheet" one to two days before the exam with numbered blank boxes. On exam days, the students filled out the Answer Sheet. The online cohort was given 10-15 minutes after the exam ended to scan and upload their work. Students were familiar with this process and had experience uploading documents prior to examination. Students used various scanning methods (e.g., smart phone apps or desktop scanner). The face-to-face cohort had their exams scanned by the instructor (ABC) and uploaded to Gradescope. A grading rubric was created by the instructor. ABC assigned graduate teaching assistants to specific questions, and they graded that question for the entire class.

In the fall 2020 semester, we surveyed students to gain a better understanding of their decisions regarding the choice of testing mode, and to quantify other affective traits about the students that may differentiate the two groups. We measured intrinsic goal orientation (Pintrich, 1993) and perceived value of chemistry (Pintrich, 1993) using a 7-point Likert scale (e.g. strongly agree to strongly disagree; **Table S1 in Appendix**). We lightly modified scales and validated them on our student population through confirmatory factor analysis. To measure engagement, we also asked students for the percent of class time they felt intellectually engaged in learning the material, with options including less than 10%; 10-30%; 31-50%; 51-70%; over 70%. To collect survey responses, we administered a Qualtrics survey to students in the class during the last week of the semester, and offered a point in extra credit (a small fraction of their total course grade, awarded for clicking on the link to the survey).

Statistical Analyses

Incoming Preparation

We obtained measures of incoming preparation through the University's Office of Institutional Research. Specifically, we obtained students' high school GPA, cumulative ACT, and SAT scores. Because the institution accepts either score, and the majority of students submit ACT scores for admissions purposes, we transformed SAT scores into the ACT scale for those students who only submitted an SAT score using the ACT.org SAT Concordance Table. As high school GPA and ACT score are highly correlated variables (r^2= 0.46, $p < 0.0001$), we developed a single measure of incoming preparation for subsequent quantitative analyses. A Principle Component Analysis (PCA) found that 73.06% of variation was explained by the first principle component. The Principle Component including high school GPA and ACT score (PC1), on which loadings were equally strong at 0.71, was then termed "Incoming Preparation" in statistical analyses.

Construct Validation

Data was often skewed (non-normal), and we used a Maximum Likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic ("MLM" estimation) in R with the lavaan package to run Confirmatory Factor Analysis (Curran *et al.*, 1996) on Intrinsic Goal Orientation and Value of Chemistry constructs. The resulting fit indices (**Table 2**) show the items belonging to a corresponding common latent factor was as intended. It should be noted that the RMSEA value for Intrinsic Goal Orientation was ranked as "mediocre". In this study with sample sizes of <200, the chi-squared test is a reasonable measure of fit, but may lead to artificial inflation of the RMSEA value. In combination with acceptable CFI, chi-squared, and SRMR values, we chose to proceed with this latent factor in analysis. We extracted factor scores for each construct from the CFA and the resulting factor score was used in all further analyses as a single representation of intrinsic goal orientation and perceived value of chemistry.

Table 2: CFA Analysis Fit Indices. Fit index measures are reported along with a description and explanation of literature supported cutoff values (Ballen and Salehi, 2021). The number of survey items with the construct are indicated with a superscript, and samples sizes are reported for each construct.

Fit Index	What is measured	Explanation	Intrinsic Goal Orientation[4] (N= 198)	Value of Chemistry[6] (N= 198)
χ^2	Determines the magnitude of discrepancy between the covariance matrix estimated by the model and the observed covariance matrix of the data sets.	Should be non-significant, meaning the estimated covariates are not significantly different from the actual data covariates	5.355 (p =0.06)	12.238 (p =0.2)
CFI	Determines if the model fits the data by comparing the χ^2 of the model with the χ^2 of the null model Adjusts for sample size and number of variables	>0.90 = Acceptable >0.95 = Good	0.964	0.993
RMSEA	Determines how well the model fit the data, and favor parsimony and a model with fewer parameters.	<0.05 to 0.06 = good 0.06 to 0.08 = acceptable 0.08 to 0.10 = mediocre >0.10 = unacceptable	0.09	0.043
SRMR	A standardized square-root of the difference between the observed correlation and the predicted correlation	< 0.05 = good 0.05 to 0.08 = acceptable 0.08 to 0.10 = mediocre >0.10 = unacceptable	0.03	0.029

Comparative Models

Statistical analyses were performed to assess differences in Intrinsic Goal Orientation, Perceived Value of Chemistry, Performance, and Engagement among students choosing between the online and face to face testing modalities. In order to test for the impact of testing modality on the latent variables extracted from CFA analysis (Intrinsic Goal Orientation and Perceived Value of Chemistry) and incoming preparation measures, the following models were utilized:

$$\text{Intrinsic Goal Orientation} = \beta_0 + \beta_1 \text{Testing Modality} + \varepsilon$$
$$\text{Perceived Value of Chemistry} = \beta_0 + \beta_1 \text{Testing Modality} + \varepsilon$$
$$\text{Incoming Preparation} = \beta_0 + \beta_1 \text{Testing Modality} + \varepsilon$$

Next, the impact of testing modality, exam number, and each of these latent were examined to explore their potential impact on student performance using the following model:

$$\text{Performance} = \beta_0 + \beta_1 \text{Testing Modality} + \beta_2 \text{Exam Number} + \beta_3 \text{Intrinsic Goal Orientation} + \beta_4 \text{Perceived Value of Chemistry} + \varepsilon$$

Lastly, to explore the effect of engagement on student performance, data were first subsetted by engagement level (<10%, 10-30%, 31-50%, 51-70%, 71-100%). Within each engagement category, one linear model was used to test for differences in student performance due to testing modality. Please note, only pairwise analyses were run within each engagement category, and not longitudinally across categories due to constrained sample size. For example, the model for reported engagement category of <10% would be:

$$\text{Engagement } <10\% = \beta_0 + \beta_1 \text{Testing Modality} + \varepsilon$$

All statistical analyses were performed in R version 4.0.3. For quantitative analysis of RQ1-RQ3, we ran repeated measures linear mixed-effect (LME) models using the nlme package (Pinheiro, Jose *et al.*, 2020). To account for repeated measures from a single student, we included Student ID as a random effect variable. In measures of performance, we also included incoming preparation as a random effect as it significantly impacted performance outcomes ($F_{(1,173)}=57.823$, $p < 0.0001$).

When appropriate, we used the emmeans package (Lenth, Russel, 2019) to obtain post-hoc pairwise significance, utilizing Tukey Post-Hoc p-value adjustments. All independent correlational measures are based on Pearson's Correlation Coefficients. Statistical significance was based on $p < 0.05$ and confidence intervals that exclude zero.

Qualitative analysis

The open-ended survey question, central to this research, gauged students' preference for taking the exams either online or face-to-face. We provided students with the option to answer one of the following two questions: (1) If you completed most exams IN LECTURE (i.e., face-to-face): Why did you choose to take exams face-to-face, rather than online? or (2) If you completed most exams ONLINE: Why did you choose to take exams ONLINE, rather than face-to-face? For both datasets, an author (AE) and a graduate research assistant created categories using open-ended coding, and for the students who completed most of the exams face-to-face, the following nine

themes emerged from their open-ended responses: (1) Preference for classroom environment, (2) Avoidance of technical issues, (3) Easy and efficiency of preparation, (4) Increased performance and focus, (5) Instructor interactions, (6) Preference for physical copy of exam, (7) Increased comfort and decreased stress, (8) Instructor recommendation, and (9) Avoidance of cheating. We coded student responses into these categories, and any responses that did not fit into a category or were not meaningful were left uncoded. For the students who completed most of the exams online, the following six themes emerged from their open-ended responses: (1) More accustomed to the online environment, (2) Increased preparation time, (3) Increased convenience, (4) Increased rest, (5) Avoidance of COVID risk factors, and (6) Decreased test anxiety and classroom stress. An author (AE) and a graduate research assistant coded each dataset separately and met weekly via Zoom to discuss any disparities until 100% agreement was met for both datasets.

An individual student response was coded into multiple themes when appropriate. In other words, a single student's response may fit into multiple thematic codes. We calculated the frequency of response within each theme and separate testing modalities by dividing the number of responses for a specific category and dividing it by the total data points gathered for one modality. This was then repeated for the second testing modality.

Results & Discussion

To our knowledge, this research is the first to explore student preferences for online or face-to-face testing modes in undergraduate chemistry. The scarcity of work on this topic is due, in part, to the relatively recent and widespread reliance on online exams resulting from the transition online during the COVID-19 pandemic. However, enrollment in online courses is quickly increasing across the United States due to their accessibility, flexibility, and convenience (Allen and Seaman, 2014).

In the broader literature, the effect of testing mode on student outcomes is mixed, with some studies showing nonsignificant differences between computer-based and paper-based test results (Horkay *et al.*, 2006; Wang *et al.*, 2007, 2008; Tsai and Shin, 2013; Meyer *et al.*, 2016) and other studies showing significant differences (Clariana and Wallace, 2002; Bennett *et al.*, 2008; Keng *et al.*, 2008; Backes and Cowan, 2019). In the following results and discussion, we summarize our findings, place them in the context of previous work, and when possible, make recommendations for future research or instructional practices.

(RQ1) Testing mode performance gaps. Student performance varied both by exam number ($F_{(2,352)}=166.072$, $p < 0.001$) and testing mode ($F_{(1,352)}=57.942$, $p < 0.001$; **Fig. 1A**). Specifically, students who chose to take their exams face-to-face outperformed their online peers overall (**Fig. 1A**), and on each individual exam (**Fig. 1B**). While average performance decreased over time, this relationship is independent of testing mode, as indicated by a non-significant interaction between testing mode and exam number ($p = 0.37$).

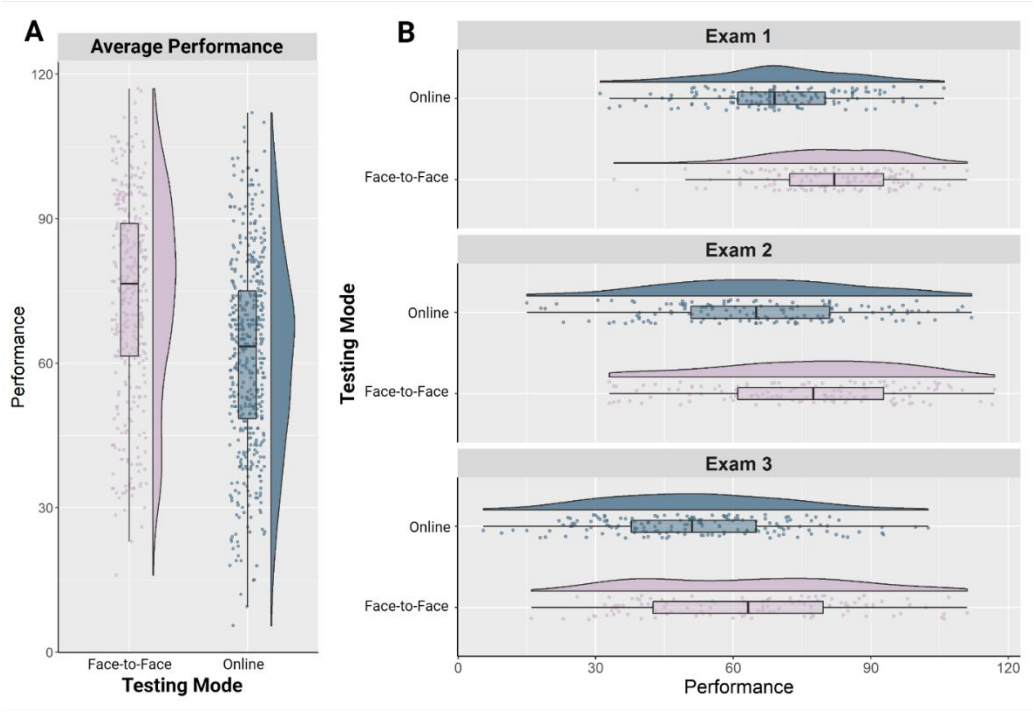

Figure 1: Exam performance outcomes by testing mode across a semester of organic chemistry. A. Average combined exam performance by testing mode. B. Average performance outcomes by testing mode across individual exams in a semester.

(RQ2) Intrinsic goal orientation, perceived value of chemistry, and incoming academic preparation. We found that testing mode was not impacted by intrinsic goal orientation ($F_{(1,394)}$=0.042, p = 0.839) or incoming preparation ($F_{(1,360)}$=0.522, p = 0.470) (**Table S2 in Appendix**). At the onset of this study, we hypothesized that incoming academic preparation would be a central factor distinguishing online testers and face-to-face testers. Incoming preparation is frequently identified as the culprit explaining performance gaps, particularly in introductory or lower division science courses (Salehi, Burkholder, *et al.*, 2019; Salehi *et al.*, 2020). After all, students who attend high schools with less academic resources are less prepared for higher education (Ferguson *et al.*, 2007, Mueller, 2007; Aikens and Barbarin, 2008) and less likely to enter higher education altogether (Sewell and Shah, 1967). While we observed a relationship between incoming preparation and performance, we did not observe a relationship between testing mode and incoming preparation that would explain performance gaps between online and face-to-face examinations.

Contrary to our predictions, the only difference between students of the two testing modes, other than test performance outcomes, was their responses to survey questions that gauged chemistry task value. The relationship between perceived value of chemistry and testing mode was statistically significant ($F_{(1,394)}$=4.393, p = 0.037), such that students who chose to take the exam in a face-to-face format reported a higher value of chemistry (**Fig. 2B**). Task value is the perceived value attributed to a task (in this case chemistry) or the reported utility and importance of the disciplinary content. Students' motivation to learn and perform may be dependent upon, in part, the value they attribute to the task, and previous work has demonstrated its predictive relation to performance

outcomes (Bong, 2001; Joo *et al.*, 2013; Robinson *et al.*, 2019). One explanation for our results may be that students who chose to take face-to-face exams did so, to some extent, based on how important they perceived organic chemistry, which in turn impacted their performance on assessments.

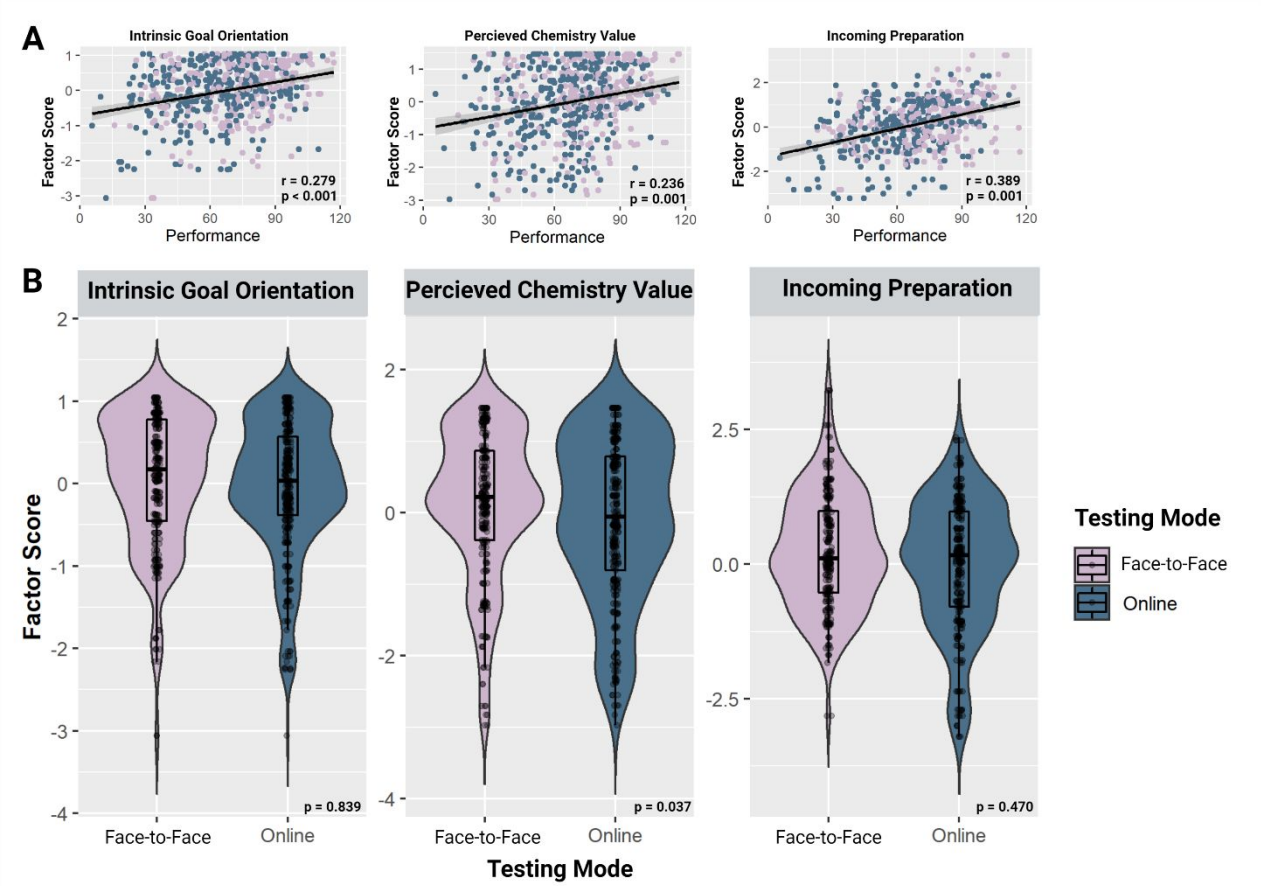

Figure 2: Impacts of affective factors (intrinsic goal orientation, perceived chemistry value, and incoming preparation) on performance outcomes across testing mode. A. Correlations between affective factors and performance. B. Affective factors by testing mode. Significant relationships include p-values.

However, we observed a positive correlation between performance outcomes and these affective measures as well as incoming academic preparation. Specifically, we found that intrinsic goal orientation ($F_{(1,173)}$=43.36, $p < 0.001$), perceived value of chemistry ($F_{(1,173)}$=10.23, $p = 0.001$), and incoming preparation ($F_{(1,173)}$=57.82, $p < 0.001$) significantly impacted student performance regardless of testing mode. When we ran correlational analyses of each measure independently, we found each measure was positively related to student performance (Intrinsic goal orientation: $r = 0.28$, $p < 0.001$; Perceived chemistry values: $r = 0.23$, $p < 0.001$; Incoming preparation: $r = 0.38$, $p < 0.001$; **Fig. 2A**). In other words, these factors correlated with academic performance for all students, but not based on testing mode preference.

(RQ3) Engagement. Descriptive statistics revealed little variance in reported engagement between the two testing modes (**Fig. 3A**) in most cases, although it does appear that student who chose online testing formats reported extremely low levels of engagement (<10%) at nearly double the rate of

face-to-face test takers. However, we did observe an overall effect of engagement on performance ($F_{(4,187)} = 6.558$, $p < 0.001$) **(Fig. 3B)**. Interestingly, when the effects of engagement on performance were analyzed by testing mode, a clear pattern arose. Despite reporting equivalent levels of in-class engagement, students who took the exam face-to-face outperformed their online peers. While this relationship is not statistically significant in students that report 0-10% and 31-50% percent engagement, this relationship was significantly significant among students who reported engaging in class 10-30% of the time ($F_{(1,66)}=6.69$, $p= 0.012$), 51-71% of the time ($F_{(1,119)}=20.03$, $p< 0.001$), and 71-100% of the time ($F_{(1,63)}=8.01$, $p= 0.006$) **(Fig. 3B)**.

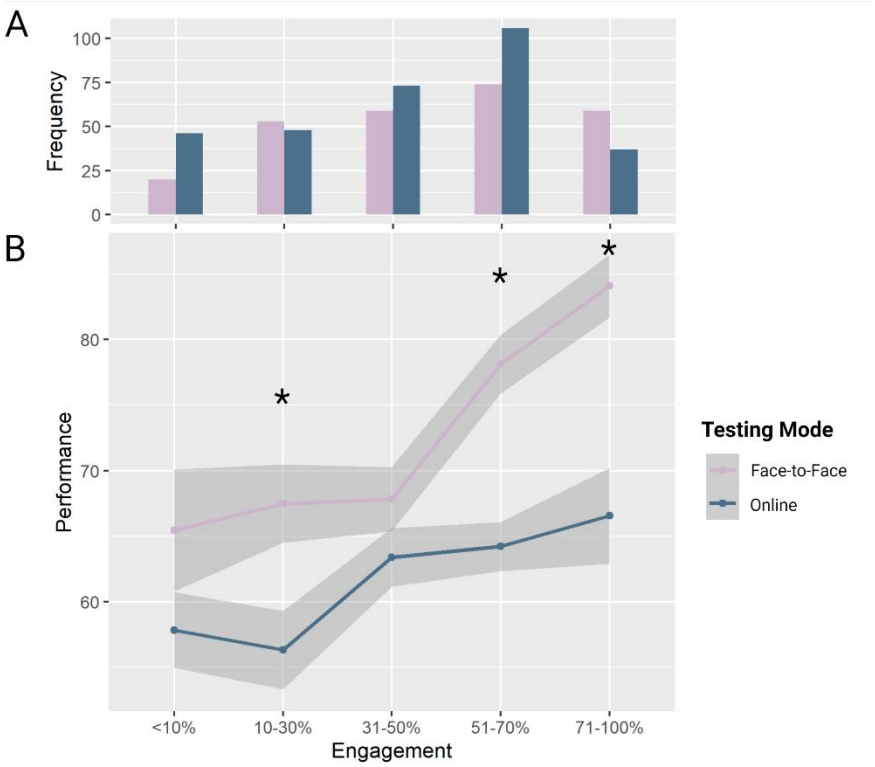

Figure 3: Relationship between reported in-class engagement and testing mode. A. Frequency of different levels of engagement show little differences between testing modes. B. Relationship between exam performance and engagement levels by testing mode. *$p < 0.05$; ***$p < 0.001$. Statistical analysis includes independent pairwise testing between formats at each engagement category.

In other words, among students who reported being intellectually engaged in learning the material for over 50% of class time, those who chose to take the exams face-to-face performed significantly better than students who reported equivalent engagement levels but took the exam online. We expected student engagement and measures of performance to be closely linked, regardless of testing mode (Coates, 2005). We suggest one of three explanations of our results. In one scenario, despite similar levels of reported engagement, other affective factors inherent to the student lead to underperformance during online exams (such as reported value of the material, as described above, or an unexplored variable). Another possibility may be that online exams directly disadvantage students who are otherwise equally prepared and engaged. Despite the importance of student engagement in evidence-based teaching and learning (National Research Council, 2012), it is critical that assessments of students are reflecting the content students have learned. And assessing students in two different ways (computer-based or paper-based) in two different locations (remote or in the

classroom) may result in the appearance of lower understanding of material, when in fact online students experience the assessment differently, leading to lower scores despite similar knowledge. Previous researchers pointed out that factors such as screen size, font size, and resolution of graphics have the potential to enhance the experience of taking online assessments (McKee and Levinson, 1990). While our results do not support this previous work, we agree that the experience of taking an online exam is fundamentally different from a face-to-face exam, which may have led to lower grades among our online testers. A third possibility is that our single-item measure of engagement, in addition to low sample sizes across some engagement categories, is not sufficient to draw conclusions at this stage; yet, we point to the possibility of a relationship between these testing modes and engagement, and hope future work pursues this open question.

(RQ4) Student preference for testing mode. We probed student preference for testing modality and observed sizable qualitative differences among student responses that supported differential preferences in the exam experience. We received 199 surveys from students who reported taking most exams online. After removing 98 surveys from students who left the open-ended response blank or who did not provide a meaningful response, we had a total of 101 responses, which we binned into 6 categories, leading to a total of 124 data points. The four most common responses (**Fig. 4**) for the online data were: Avoidance of COVID risk factors (38%), Increased convenience (37%), More accustomed to the online environment (8%), and Decreased test anxiety and classroom stress (8%).

We received 199 surveys from students who reported taking most exams face-to-face. After removing 104 surveys due to non-response, we had a total of 95 responses, which we binned into 9 categories, leading to a total of 167 data points. The four most common responses for the face-to-face data (**Fig. 4**) were: Increased performance and focus (20%), Increased comfort and decreased stress (17%), Preference for classroom environment (17%), and Avoidance of technical issues (16%).

Regardless of whether a student chose online or face-to-face testing modes, they were likely to mention comfort and convenience as a reason for their preference, whether that is through decreased anxiety, exam format, or preference for a specific environment. While nearly all the top responses for online testing modalities are related to comfort, with the notable exception of COVID-19 risk factors, students who chose face-to-face modalities mention factors of convenience *and* classroom success. Bringing into question: how do students identify what is "convenient, comfortable, and important" within a learning environment? For example, future work should focus on understanding how students classify convenience and comfort to illuminate why students choose their preferred testing modalities, and to determine if underlying personality traits inherent to those decisions play a role in the performance disparity.

	Theme	Frequency	Description	Example
Face to Face Testing Modality	Increased Performance and Focus	19.3%	Students reference an absence of distractions or an increased ability to focus	"less distractions"
	Preference for Classroom Environment	17.5%	Students reference suitability of "normal" atmosphere, importance of classmates, and avoidance of home environments	"the atmosphere and experience is as close to normal as possible"
	Increased Comfort and Decreased Stress	17.5%	Students indicate increased comfort and decreased stress due to external factors and internal factors.	"more used to in person exams" "online exams give me stress"
	Avoidance of Technological Issues	16.3%	Students reference the concerns of internet and technological complications (ie. Canvas)	"bad internet" "error and complication"
	Ease and Efficiency of Preparation	10.2%	Students indicate a decreased risk of issues turning in assessments on time, or simplicity of turning in assessments.	"don't have to worry about submitting"
	Preference for Physical Copy of Exam	8.4%	Students reference a distaste for screen time, a preference for paper-exams, or the ability to take notes on the physical test	"prefer paper over screen" "like to be able to mark through things"
	Instructor Recommendation	6.0%	Students reference the instructor recommendation for in-person exams, or statistics indicated increased performance in face-to-face environments	"Statistics showed students do much better"
	Instructor Interactions	2.4%	Students reference the ability to ask the instructor questions during the exam	"ask questions directly" "ask and get a clearer answer"
	Avoidance of Cheating	2.4%	Students reference a fear of being falsely accused of cheating or increased exam integrity	"nervous about being accused" "integrity of my exam"
Online Testing Modality	Increased Convenience	37.7%	Students reference the convenience and accessibility of the online format	"hard to get through traffic" "had another zoom quickly after class"
	Avoidance of COVID Risk Factors	36.9%	Students reference personal quarantines, risks of COVID infection, or the avoidance of other students	"more comfortable because don't have to be in contact with people" "safer online"
	Decreased Test Anxiety and Classroom Stress	8.2%	Students reference decreased anxiety due to both internal and external factors	"being in a room with people worsens it"
	More Accustomed to the Online Environment	8.2%	Students reference the need for consistency of learning and test-taking environments or that they had grown accustomed to online exams	"learned the material online so wanted to be tested online" "get used to that format"
	Increased Preparation Time	6.6%	Students reference time gained due to lack of transportation time or studying immediately prior to opening online exams	"more time to study"
	Increased Rest	2.4%	Students reference more resting time	"didn't have to get up as early"

Figure 4: Themes resulting from coded open-ended survey items including the frequency of response, a description of the theme, and examples extracted from student responses.

Many students across both testing modes mentioned stress or anxiety associated with the exams in their open responses. Test anxiety can be characterized by the negative cognitive or emotional reactions to perceived and actual stress from fear of failure (Zeidner, 1998) Test anxiety is pervasive across large foundational STEM courses such as organic chemistry, where exam grades account for the majority of the student's final score, and previous work shows this disproportionately impacts women (Ballen *et al.*, 2017; Salehi, Cotner, *et al.*, 2019). To our knowledge, research has not addressed test anxiety during online exams, or compared the relative impact of online and face-to-face exams on anxiety, but this is a potential area of future exploration.

Synthesis

In this exploratory study, we investigated several factors that may be associated with differential performance outcomes among online and face-to-face testers in organic chemistry. While we found no evidence that differences in *engagement*, *intrinsic goal orientation* or *incoming preparation* was associated with this relationship, it does appear that differences could be related to increased *value of chemistry* among those who opted for face-to-face examinations, as they displayed higher levels of task value. On a numeric scale, these students rated that the course content was important for them to know, that they will be able to use the content in later studies, and that they enjoyed learning the content at higher levels than online students (Pintrich et al., 1993). Future interventions in organic chemistry can target students with low task value by contextualizing course materials in meaningful way (Fahlman *et al.*, 2015), potentially closing the observed performance gap displayed here.

Differential performance may be due to alternative explanations only vaguely explored within this study. Differential performance may be explained by (1) measurable advantages of the in-person classroom testing environment unnoticed and unreported by students in this study, or (2) unmeasured advantages of ritualistic behaviors exhibited by students traveling to a classroom for in-person examination. Previous research has shown that ritualistic behaviors such as test-taking routines and factors such as the use of professional attire can increase student performance (Adam and Galinsky, 2012). This can be due to the increased perception of professional expectations by students who chose to dress professionally, or through the introduction of a routine specific to test taking. In another example, previous work showed that chewing gum for up to 5 minutes before an examination may increase student performance (Onyper *et al.*, 2011). Future work will profit from exploring the ritualistic act of relocating from a home environment to a classroom environment as a key part of the testing preparation routine, transitioning to a test-taking mindset, and establishing boundaries between a relaxed state and a professional state for increased test performance.

Limitations and Future Directions

Limitations

Like many discipline-based research studies, our study relied on student self-reports as the primary data source, which although informative, may still fall short in obtaining unbiased responses to survey questions. Because we documented student data at the end of the semester, students must recall how they performed retroactively rather than in the moment. Additionally, we are unable to identify with confidence *why* students who took exams online underperformed relative to face-to-face students.

As this study was completed in an exploratory nature, there are additional potential influences on student outcome. For example, while the instructor has extensive experience designing and teaching chemistry online, online exam opportunities for organic chemistry was a newly offered option. By design, both exam formats required students to complete identical answer sheets and online students then uploaded responses to be graded. Students who took the test on the online format had previous experience uploading activity sheets and were familiar with the process, denoting that neither testing format exposed students to new challenges on exam day. While there were no

reportable issues as the online option continued through the semester and there are no planned adaptations to implementation in the future, it is possible that future iterations could lead to changes in student outcomes and perceptions following adjustments.

Future research will delve into whether these results were due to the use of a computer or the physical space in which students took exams, and how these experiences led to lower performance and value they placed on chemistry as a field. However, taken together, our results were consistent and clear, and reflect the divergent experiences of students who must decide how they approach assessments. We hope our research can serve as a foundation for future questions that tease apart impacts of task value and testing mode preference on student performance.

Conclusion

We found that students who elected to take online tests underperformed relative to those who took face-to-face tests across the semester, and a significant difference between these two groups of students was how valuable they perceived chemistry as a discipline, as well as their open-ended responses detailing their personal motivations in taking online or face-to-face exams. Surprisingly, our data suggested that this relationship was not associated with *incoming preparation*, student reported *engagement* levels, or measures of *intrinsic goal orientation*. Our exploratory results support that a primary difference between test mode preferences is *perceived value of chemistry*. However, it is also possible that the relationship is due to innate aspects of classroom environment unrealized by students, or the mentality of test taking itself.

Acknowledgements

We are grateful to Tashitso Anamza for assistance with coding; the biology education research group and the discipline-based education research group at Auburn University for valuable conversations about online assessments, especially: Emily Driessen, Sharday Ewell, Chloe Josefson, Todd Lamb, and Ash Zemenick. We appreciate the undergraduate learning assistants and graduate teaching assistants who played vital roles in the course, and the undergraduates in organic chemistry who were willing to participate in our study. This work was supported by NSF DUE-2120934 awarded AEB and CJB; DUE-2011995 awarded to CJB. Figures were created with BioRender.com. This research was determined to be exempt from Auburn University's Institutional Review Board, and approved with the use of incentive for participation (Protocol #21-320 EX 2107).

Appendix

The appendix presents a copy of supplemental tables 1 and 2.

Table S1. Items used in a survey of undergraduate organic chemistry students. Students responded to survey items on a 7-point Likert scale unless otherwise noted.

Intrinsic Goal Orientation Construct (modified from Pintrich et al., 1993)
I prefer courses that are challenging so I can learn new things
I work on practice exercises and answer end of chapter questions even when I don't have to
I work hard to get a good grade even when I don't like a course
Even when study materials are dull and uninteresting, I keep working until I finish
I often choose to write about topics I will learn something from even if they require more work
Task Value (modified from Pintrich et al., 1993)
It is important for me to learn what is being taught in this course
I like what I am learning in this course
I think I will be able to use what I learn in this course in later studies
I am very interested in the content area of this course
I think that what I am learning in this course is useful for me to know
Understanding this subject is important to me
Incoming Preparation
Principal component including high school GPA and ACT score (see main text for more information)
Engagement
Percent of class time I felt intellectually engaged in learning the material: less than 10%; 10-30%; 31-50%; 51-70%; over 70%

Table S2. Sample size, mean, standard deviation, 95% confidence interval, skewness, and kurtosis measures for latent variables and student incoming preparation measures.

	Format	*N*	*Mean*	*SD*	*CI*	*Skewness*	*Kurtosis*
Intrinsic Goal Orientation	Face to Face	275	0.0592	0.801	0.095	-0.999	0.962
	Online	318	-0.0477	0.810	0.089	-0.893	0.660
Perceived Value of Chemistry	Face to Face	272	0.1173	1.031	0.123	-0.869	0.371
	Online	321	-0.0980	1.130	0.124	-0.518	-0.585
Incoming Preparation	Face to Face	256	0.2414	1.064	0.131	0.147	-0.081
	Online	286	-0.0230	1.223	0.142	-0.578	-0.094

References

Adam H. and Galinsky A. D., (2012), Enclothed cognition. *Journal of Experimental Social Psychology*, **48**(4), 918–925.

Aikens N. L. and Barbarin O., (2008), Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology*, **100**(2), 235–251.

Allen I. E. and Seaman J., (2014), Grade Change: Tracking Online Education in the United States. *Babson Survey Research Group*.

Backes B. and Cowan J., (2019), Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review*, **68**, 89–103.

Ballen C. J. and Salehi S., (2021), Mediation Analysis in Discipline-Based Education Research Using Structural Equation Modeling: Beyond "What Works" to Understand How It Works, and for Whom. *J Microbiol Biol Educ*.

Ballen C. J., Salehi S., and Cotner S., (2017), Exams disadvantage women in introductory biology. *PLOS ONE*, **12**(10), e0186419.

Barr D. A., Gonzalez M. E., and Wanat S. F., (2008), The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students. *Academic Medicine*, **83**(5), 503–511.

Bennett R. E., Braswell J., Oranje A., Sandene B., Kaplan B., and Yan F., (2008), Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, **6**(9).

Black A. E. and Deci E. L., (2000), The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science education*, **84**(6), 740–756.

Bong M., (2001), Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary educational psychology*, **26**(4), 553–570.

Chi M. T. H. and Wylie R., (2014), The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, **49**(4), 219–243.

Clariana R. and Wallace P., (2002), Paper–based versus computer–based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, **33**(5), 593–602.

Coates H., (2005), The value of student engagement for higher education quality assurance. *Quality in higher education*, **11**(1), 25–36.

Curran P. J., West S., and Finch J. F., (1996), The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis. *Psychological methods*, **1**(1), 16–29.

Eddy S. L., Converse M., and Wenderoth M. P., (2015), PORTAAL: A Classroom Observation Tool Assessing Evidence-Based Teaching Practices for Active Learning in Large Science, Technology, Engineering, and Mathematics Classes. *LSE*, **14**(2), ar23.

Fahlman B. D., Purvis-Roberts K. L., Kirk J. S., Bentley A. K., Daubenmire P. L., Ellis J. P., and Mury M. T., (2015), *Chemistry in context: applying chemistry to society*, McGraw-Hill,.

Fautch J. M., (2015), The flipped classroom for teaching organic chemistry in small classes: is it effective? *Chem. Educ. Res. Pract.*, **16**(1), 179–186.

Ferguson H., Bovaird S., and Mueller M., (2007), The impact of poverty on educational outcomes for children. *Paediatr Child Health*, **12**(8), 701–706.

Ferrell B., Phillips M. M., and Barbera J., (2016), Connecting achievement motivation to performance in general chemistry. *Chemistry Education Research and Practice*, **17**(4), 1054–1066.

Garcia T., (1993), Women and Minorities in Science: Motivational and Cognitive Correlates of Achievement.

Hochlehnert A., Brass K., Moeltner A., and Juenger J., (2011), Does medical students' preference of test format (computer-based vs. paper-based) have an influence on performance? *BMC medical education*, **11**(1), 1–6.

Horkay N., Bennett R. E., Allen N., Kaplan B., and Yan F., (2006), Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, **5**(2), n2.

Hussar B., Zhang J., Hein S., Wang K., Roberts A., Cui J., et al., The Condition of Education 2020. 348.

Joo Y. J., Lim K. Y., and Kim J., (2013), Locus of control, self-efficacy, and task value as predictors of learning outcome in an online university context. *Computers & Education*, **62**, 149–158.

Keng L., McClarty K. L., and Davis L. L., (2008), Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education*, **21**(3), 207–226.

Lane E. S. and Harris S. E., (2015), A New Tool for Measuring Student Behavioral Engagement in Large University Classes. *Journal of College Science Teaching*, **44**(6), 83–91.

Lenth, Russel, (2019), *emmeans: Estimated Marginal Means, aka Least-Squares Means*,.

McKee L. M. and Levinson E. M., (1990), A review of the computerized version of the Self-Directed Search. *The Career Development Quarterly*, **38**(4), 325–333.

McNeal K. S., Zhong M., Soltis N. A., Doukopoulos L., Johnson E. T., Courtney S., et al., (2020), Biosensors show promise as a measure of student engagement in a large introductory biology course. *CBE—Life Sciences Education*, **19**(4), ar50.

Mervis J., (2011), Weed-out courses hamper diversity. *Science*, **334**(6061), 1333.

Meyer A. J., Innes S. I., Stomski N. J., and Armson A. J., (2016), Student performance on practical gross anatomy examinations is not affected by assessment modality. *Anatomical sciences education*, **9**(2), 111–120.

Miltiadous A., Callahan D. L., and Schultz M., (2020), Exploring engagement as a predictor of success in the transition to online learning in first year chemistry. *Journal of Chemical Education*, **97**(9), 2494–2501.

National Research Council, (2012), *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, Singer S. R., Nielsen N. R., and Schweingruber H. A. (eds.) The National Academies Press.

Onyper S. V., Carr T. L., Farrar J. S., and Floyd B. R., (2011), Cognitive advantages of chewing gum. Now you see them, now you don't. *Appetite*, **57**(2), 321–328.

Ost B., (2010), The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, **29**(6), 923–934.

Pinheiro, Jose, Bates, Douglas, DebRoy, Saikat, Sarkar, Deepayar, and R Core Team, (2020), *nlme: Linear and Nonlinear Mixed Effects Models*,.

Pintrich P. R., (1999), The role of motivation in promoting and sustaining self-regulated learning. *International journal of educational research*, **31**(6), 459–470.

Pintrich P. R., Smith D. A. F., Garcia T., and Mckeachie W. J., (1993), Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement*, **53**(3), 801–813.

Prisacari A. A. and Danielson J., (2017), Rethinking testing mode: Should I offer my next chemistry test on paper or computer? *Computers & Education*, **106**, 1–12.

Pritchard G. M., Rules of Engagement: How Students Engage With Their Studies.

Rask K., (2010), Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, **29**(6), 892–900.

Robinson K. A., Lee Y., Bovee E. A., Perez T., Walton S. P., Briedis D., and Linnenbrink-Garcia L., (2019), Motivation in transition: Development and roles of expectancy, task values, and costs in early college engineering. *Journal of Educational Psychology*, **111**(6), 1081.

Salehi S., Burkholder E., Lepage G. P., Pollock S., and Wieman C., (2019), Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Phys. Rev. Phys. Educ. Res.*, **15**(2), 020114.

Salehi S., Cotner S., Azarin S. M., Carlson E. E., Driessen M., Ferry V. E., et al., (2019), Gender Performance Gaps Across Different Assessment Methods and the Underlying Mechanisms: The Case of Incoming Preparation and Test Anxiety. *Frontiers in Education*, **4**, 107.

Salehi S., Cotner S., and Ballen C. J., (2020), Variation in Incoming Academic Preparation: Consequences for Minority and First-Generation Students. *Frontiers in Education*, **5**, 170.

Sawada D., Piburn M. D., Judson E., Turley J., Falconer K., Benford R., and Bloom I., (2002), Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, **102**(6), 245–253.

Sewell W. H. and Shah V. P., (1967), Socioeconomic Status, Intelligence, and the Attainment of Higher Education. *Sociology of Education*, **40**(1), 1–23.

Seymour E. and Hunter A.-B., (2019), Talking about leaving revisited. *Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education*.

Smith M. K., Jones F. H. M., Gilbert S. L., and Wieman C. E., (2013), The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE Life Sci Educ*, **12**(4), 618–627.

Test Anxiety:: The State of the Art (Perspectives on Individual Differences) 1st Edition by Zeidner, Moshe published by Springer, (1998), Springer.

Tsai T.-H. and Shin C. D., (2013), A Score Comparability Study for the NBDHE: Paper–Pencil Versus Computer Versions. *Evaluation & the Health Professions*, **36**(2), 228–239.

Wang S., Jiao H., Young M. J., Brooks T., and Olson J., (2007), A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, **67**(2), 219–238.

Wang S., Jiao H., Young M. J., Brooks T., and Olson J., (2008), Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and psychological measurement*, **68**(1), 5–24.

Wiggins B. L., Eddy S. L., Wener-Fligner L., Freisem K., Grunspan D. Z., Theobald E. J., et al., (2017), ASPECT: A Survey to Assess Student Perspective of Engagement in an Active-Learning Classroom. *LSE*, **16**(2), ar32.