Journal of the Scholarship of Teaching and Learning, Vol. 24, No. 2, June 2024, pp.79-89.
doi: 10.14434/josotl.v24i2.35209

Deconstructing the Testing Mode Effect: Analyzing the Difference Between Writing and No Writing on the Test

Daniel M. Settlage
University of Arkansas-Fort Smith
Dan.Settlage@uafs.edu

Jim R. Wollscheid
University of Arkansas-Fort Smith
Jim.Wollscheid@uafs.edu

Abstract: The examination of the testing mode effect has received increased attention as higher education has shifted to remote testing during the COVID-19 pandemic. We believe the testing mode effect consists of four components: the ability to physically write on the test, the method of answer recording, the proctoring/testing environment, and the effect testing mode has on instructor question selection. This paper examines the first component, the ability to write on the test, which we believe is a neglected area of study. Using a normalization technique to control for student aptitude and instructor bias, we find that removing the ability of students to physically write on the test significantly lowers student performance. This finding holds across multiple question types classified by difficulty level, Bloom's taxonomy, and on figure/graph-based questions, and has implications for testing in both face-to-face and online environments.
Keywords: testing mode, test environment, writing on the test, paper versus computer testing, test construction

Over the past 20 years in higher education, the mode in which students take their course exams has broadened considerably. In addition to traditional paper exams, computerized exams (both in-person and remote) have become increasingly prevalent. The COVID-19 pandemic accelerated this shift to varying testing modalities. As institutions increasingly shifted their face-to-face and hybrid classes to online formats, a concomitant switch from paper exams to electronically delivered online examinations occurred. As higher education moves forward to the "new normal" in a post-pandemic environment, students will likely face a greater diversity in classroom modes of delivery and testing conditions. Clark (1994) states that good instructional design requires that different instructional techniques should produce equivalent results if the content and cognitive activities are identical. Given the implicit goal of good instructional design, it is important for faculty to understand the possible differences in student performance that may arise solely due to differences in testing conditions.

This testing mode effect refers to the difference in student performance on identical tests that were administered using different test administration methods (e.g. computer-based versus paper-based examinations). The testing mode effect has become an important research topic in a variety of disciplines. Most of the work in this arena has focused on examining the performance difference between computer-based tests and paper tests, and the literature has produced mixed results depending on the specific discipline, type, and audience of the test. Smolinsky, et al. (2020) found that students in calculus courses performed slightly better on formative assessments using paper-based tests compared to computer-based tests. Backes & Cowan (2019) discovered that K-12 students taking standardized tests in Massachusetts performed better on paper tests as compared to the computerized versions, with more pronounced benefits in English language arts compared to mathematics. In contrast, Priscari & Danielson (2017) determined no differences when they examined the performance of chemistry students on low-stakes quizzes and practice tests between computer and paper-based tests. Additional studies (e.g. Clariana & Wallace, 2002; Randall et al., 2012; Sherman et al., 2019; Sireci et al., 2012; Vispoel et al., 2019) highlighted the mixed results surrounding computer-based vs.

traditional paper tests.

The International Testing Commission (2006) stated that a test implemented across multiple modes may be considered equivalent and reliable if similar scores are obtained. Testing equivalence is important when a test is delivered across multiple modes of delivery and used to assess student learning and performance. By directly comparing paper to online testing, the current body of literature has overlooked some key factors that can influence student performance. The literature generally omits describing and deconstructing the conditions under which the test was administered in both the online and paper-based testing environment. We assert that the testing mode effect should be deconstructed into four distinct components in order to properly study and analyze the testing mode effect as it relates to comparing various testing modalities. These four separate and distinct components that have been heretofore commingled in the literature:

1. The effect that the ability to physically write on the test has on student performance.
2. The effect of the method of answer recording (Scantron vs. paper recording vs. selecting/typing on a computer).
3. The effect the proctoring/test-taking environment has on student performance (no proctoring vs. in-class proctoring by a professor vs. online proctoring done by a third party through either web cameras or artificial intelligence).
4. The effect the test-taking environment has on the questions selected/used by the instructor for the test.

Each of these factors may influence student performance on tests in different ways. To properly analyze and fully understand differences in student performance between different testing modalities, the effect of each of these changes in the testing environment should be examined. An examination of each of these effects can lead to a greater understanding of which factors enhance or degrade student performance. This will aid in effectively crafting testing strategies/conditions to mitigate the effect the testing environment has on student performance. Most of the literature examining the testing mode effect has not identified which of the components are (or are not) driving differences in student performance across testing modes. For example, Smolinsky, et al. (2020) compared pencil and paper assignments that allowed for handwritten work with partial credit to online exams with multiple choice and no written work or partial credit. Hensley (2015) compared timed paper-based quizzes that allowed student writing with timed computerized quizzes that disallowed writing. The literature would be enhanced by studies that identify the underlying characteristics that drive differences in performance based on testing conditions. Since online testing effectively forces the instructor into certain testing modalities, a better understanding of the effect of each component of the testing mode will allow instructors to design assessment instruments that are consistent and repeatable across all testing modalities used by the instructor.

This study examines the first component of the testing mode effect, namely the effect that the ability to write on the test (or the lack thereof) has on student test performance. We do this by conducting a controlled experiment whereby students are (or are not) allowed to write on the test on an in-person paper-based test. We control for underlying student characteristics by following a data normalization procedure to enable a direct head-to-head comparison between samples. The remainder of the paper is organized as follows. First, we discuss our hypotheses. Then, we present our methodology and data, followed by our results and discussion. Finally, we present our conclusions and suggestions for future work.

Hypotheses

We examine the first component of the testing mode effect by allowing (or denying) students the ability to physically write on a paper-based test. In addition to the overall effect, we seek to understand if this effect is uniform across all questions or if it disproportionately affects certain types of questions. We seek to test the following null hypotheses:

> *Hypothesis 1: Students in each cohort (writing on the test and no writing on the test) will demonstrate no significant difference in overall test performance.*
> *Hypothesis 2: Students in each cohort (writing on the test and no writing on the test) will demonstrate no significant difference in test performance on questions categorized by the publisher based on level of difficulty.*
> *Hypothesis 3: Students in each cohort (writing on the test and no writing on the test) will demonstrate no significant difference in test performance on questions categorized by the use of figures/graphs or mathematical computations.*
> *Hypothesis 4: Students in each cohort (writing on the test and no writing on the test) will demonstrate no significant difference in test performance on questions categorized by the publisher by the level of Bloom's taxonomy.*

Our hypotheses are constructed to not only examine the effect that writing (or not) has on student test performance overall but to test the effect based on the characteristics of the questions on the test. The division of our hypotheses into difficulty, question type, and Bloom's taxonomy allows the results to be applicable in a wide variety of settings and disciplines. Lower-level courses may focus more on questions in the lower echelons of Bloom's taxonomy, while upper-level courses may do the opposite. In addition, disciplines that have a heavy emphasis on figures or math may see results that differ from disciplines that focus on conceptual questions. By disaggregating our results, practitioners can focus on the types of questions they are most likely to employ in their classes.

Methodology

The analysis was conducted in three sections of a Principles of Macroeconomics course at a mid-sized regional institution located in the mid-south of the United States prior to the COVID-19 pandemic (fall 2017 and spring 2018). Students in the fall cohort were given the ability to write on the tests (*number of students = 79*) and the students in the spring cohort were not allowed to write on the tests (*number of students = 117).* The four tests given during each semester consist of three unit tests and a comprehensive final exam. The tests and time limits used were identical for both semesters, and the tests were tightly controlled to prevent cheating. The students only had access to the tests in either a classroom setting where cell phones were not permitted or under direct supervision of the professor during office hours to discuss performance on the tests. For the cohort where students were permitted to write on the test, the number of tests printed matched the number of students in the classroom. The students that took the test were required to write their names on the tests. For the cohort where students were not allowed to write on the test, the instructor numbered each test to ensure that all tests that were utilized by students were returned to the instructor. In both cohorts the students recorded their final answers on a Scantron, and they were required to write their test number on the Scantron so that the professor could match each student to a specific copy of the test. In addition, the textbook, online homework assignments and quizzes, the course instructor, lecture notes, and all related course materials used were the same for both semesters. The fall cohort serves as the treatment group, as all tests in that semester were given with the ability to write on the test, while the spring cohort serves as the control group.

Three midterm unit tests are used to test the effects of allowing students to write on the test. The fall cohort could write on all three tests while the spring cohort could not. Though the spring cohort was not allowed to write on the test, each student was given a blank sheet of scratch paper for each test. All other testing conditions remained the same (e.g. calculator usage, class time of day, number of instructional periods between tests, answers recorded via Scantron, etc.) Presumably, any computations or graphing a student would normally do by directly writing on the test could be done by writing on the scratch paper instead. Admittedly it is more "work" to use the scratch paper, as students may have to redraw graphs or tables that appear on the test, rather than just directly modifying the figures on the printed test. All the scratch paper was collected at the end of the test to prevent information about the test from leaving the classroom. To our surprise, most of the scratch paper sheets (over 90%) were turned back with little or no writing on them. After each test was turned in, the instructor examined the printed test copy to ensure that it was not written on by the student.

The final exam was given to both cohorts with the ability to write on the test. Thus, the treatment of no writing on the test was removed for the final exam. Giving the final exam under identical testing conditions to both cohorts allows us to use the final exam to normalize the results between the two samples of students, effectively controlling for student aptitude between the two groups. Normalization procedures in studies that measure student learning have been utilized in a variety of educational studies starting with Hake (1998) and in multiple subject areas including chemistry (Bretz & McClary, 2015), engineering (Abdul, et al., 2016), physics (Marshman & Singh, 2016) and economics (Settlage & Wollscheid, 2019). The method of normalization laid out by Settlage & Wollscheid (2019) is a simple method of controlling for differences in aptitude and other underlying (and unobserved/imperfectly observed) factors. This method allows for a direct head-to-head comparison of the data, thus teasing out the true effect writing (and not writing) has on student performance.

All unit tests are multiple-choice, which is a common practice in lower-level principles of economics courses. Siegfried & Kennedy (1995) found that approximately 2/3 of assessments or tests given at the principles level are done through a multiple-choice format. Katz et al. (2000) argued that utilizing multiple-choice may present difficulties in testing deeper understanding or higher levels of Bloom's taxonomy as compared to short-answer or essay questions. However, Wainer & Thissen (1993) and Walstad & Becker (1994) showed that multiple-choice testing results were highly correlated to results from essay-based questions. This finding provides support for the common practice of using multiple questions to measure student performance. Following Walstad (1998), we utilized multiple-choice tests in this study to eliminate any potential grading bias that may occur with short-answer and essay questions.

Previous studies examining student test performance based on changes in the testing environment often use an experimental setup whereby the instructor creates a test specifically for that intervention and randomly subdivides the class into groups with and without the intervention. Although sound in principle, this methodology can suffer from question selection bias as the instructor may consciously or unconsciously select questions that may be advantageous or disadvantageous to the intervention being studied. By creating the testing instrument for the express purposes of testing the effect of an intervention, the instructor can inadvertently bias their own work. The tests used in our study were written for and used in a previous semester before the intervention occurred. That is, they were constructed before the idea of studying the effect of writing on the test arose. In this way, we avoid any potential question selection bias regarding difficulty level, question type (figure or math), or Bloom's taxonomy in question selection by the instructor. One limitation of this approach is that there may not be an equal distribution of question types across all of Bloom's taxonomy levels, difficulty levels, or question types. Thus, some question classifications have more representation than others do.

Data

The study collects the performance of the students for each question from the three unit tests and a comprehensive final exam. The three unit tests contain 50 multiple choice questions each, while the final exam contains 100 multiple-choice questions. Student performance is not only measured in aggregate but is subdivided by question type (math/figure, or neither), level of difficulty, and level of complexity based on Bloom's taxonomy. The question type is determined by the authors by using simple rules to mitigate classification bias. Questions are classified as math questions if they require any mathematical calculation to solve the question, while questions are classified as figure/graph questions if they have a figure or graph preprinted on the test that is used to solve the question. Questions that may (or may not) be solved by drawing a figure or graph were not classified as figure questions unless they also had a preprinted figure or graph to accompany them.

The level of difficulty and Bloom's level of taxonomy for each question is pre-determined by an outside source (the textbook company), to alleviate potential classification bias by the researchers. The publisher classifies the level of difficulty (easy, moderate, or challenging) based on the plausibility of the distractors while the level of thought is classified using a modified version of Bloom's taxonomy as (1) Definitional, (2) Interpretive, (3) Applicative, and (4) Analytical.

Similar studies in the literature examining differences in test scores based on a treatment typically relied on simply examining the difference in the performance of the treatment and control groups. One problem with this approach is that the composition of the students in the class can be significantly different from one section to the next, which can lead to spurious conclusions regarding the effect of interventions on student learning. Following Settlage & Wollscheid (2019), our study uses a normalization procedure to correct for differences in underlying student performance and aptitude across sections. The normalization process uses the mean scores for a single test (the normalization test) that was given under essentially identical conditions in both cohorts (writing and no writing). Since the tests and testing conditions were identical between the cohorts, any difference between the scores in the two groups can be attributed to the differences in individual student characteristics and abilities (such as ACT score, GPA, previous knowledge in economics, etc.). For example, suppose the control group scores 2% better than the treatment group on the normalization test. Since the normalization test was given under identical conditions to both groups, it is reasonable to assume that the gap is due to differences in individual student characteristics in each cohort. Assuming the treatment had no effect, we would expect this gap to persist for all tests. If the gap narrows (or widens) we can attribute that change to the effect of the treatment. This method of controlling for varying underlying student attributes is simpler, more powerful, and less invasive than the traditional and labor-intensive method of gathering data on underlying student demographics. We rely on revealed student behavior to correct for underlying student attribute differences. To normalize the data, we adjust the performance data using the procedure illustrated in figure 1.

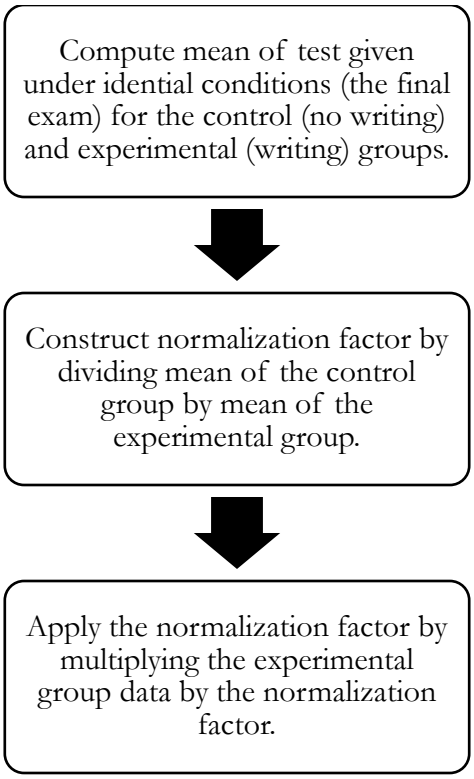

Figure 1. Normalization Procedure.

This procedure normalizes the data from the writing cohort to be equal to the data from the no writing cohort on the final exam, with the normalization factor being the mean of the no writing cohort final divided by the mean of the writing cohort final. Thus, we are standardizing the data on the final exam, which was administered to both cohorts under identical conditions whereby all students were allowed to write on the test. Using this normalization procedure, any differences detected on tests 1-3 can be attributed to the effect of writing on the test and not due to differences in student characteristics between the two cohorts. The above method controls for student aptitude and omitted variables, rather than naively assuming that both sections are of equal aptitude and have no possible omitted variable bias.

The mean on the final exam for the no writing cohort is 69.7% while the mean of the writing cohort on the same exam is 65.1%, leading to a normalization factor of 1.07. Since the final exam was given under identical conditions to both cohorts, we attribute this performance difference to differences in individual student characteristics for each cohort. When applied to tests 1-3, this normalization process allows the results of the no writing cohort and the writing cohort to be directly compared. When utilizing the normalized data, any differences between the performances of the two cohorts on tests 1-3 are directly attributable to the writing effect.

Results and Discussion

After applying the normalization process, we see that in Table 1 writing on the test shows a substantial improvement as compared to no writing on the test. There is a statistically significant difference in performance for test 2, test 3, and overall. Students allowed to write on the test improved their overall performance by 3.5%, which represents an average increase of 1.75 questions for the writing cohort. The evidence in Table 1 leads us to reject hypothesis 1 (students in each cohort will demonstrate no

significant difference in overall test performance).

Table 1. Test Performance (with Normalization).

	No Writing	Writing	Difference	p-values
Test 1 (n=50)	68.6%	70.0%	+1.48%	0.2050
Test 2 (n=50)	67.4%	74.4%	**+6.98%**	**0.0000**
Test 3 (n=50)	53.8%	55.8%	**+2.03%**	**0.0527**
Overall (n=150)	63.3%	66.8%	**+3.50%**	**0.0000**

Note. All mean performances for the writing sample have been normalized.
P-values are from the paired two-tailed t-test, with statistically significant results (at the 10% level) bolded.

Table 2 reports the difference in performance between the two cohorts broken out by question difficulty level. From Table 2, we see that students allowed to write on the test show an increase in performance compared to those who were prohibited from writing on the test. There are statistically significant improvements for questions rated as easy and moderate, while questions rated as challenging do not show a statistically significant difference. This lack of significance on questions classified as challenging may be due to the limited number of questions given at this level. The higher performance of the writing cohort on easy and moderate questions may be due to students making fewer "mistakes" on the test when transcribing their answers to a Scantron form. Prior to bubbling in an answer on a Scantron, students in the writing cohort may indicate their answer choices directly on their test by circling their answer or crossing out distractors they deem incorrect. By prohibiting students from writing on the test we prevent this behavior, which may have a deleterious effect on student performance. The evidence in Table 2 leads us to reject hypothesis 2 (students in each cohort will demonstrate no significant difference in test performance on questions categorized by the publisher based on level of difficulty) for easy and moderate questions and fail to reject hypothesis 2 for questions rated as challenging.

Table 2. Performance by Question Difficulty.

Difficulty Level	No Writing	Writing	Difference	p-values
Easy (n=49)	66.8%	71.4%	**+4.7%**	**0.0001**
Moderate (n=96)	61.8%	64.8%	**+3.0%**	**0.0006**
Challenging (n=5)	57.6%	59.7%	+2.0%	0.3970

Note. All mean performances for the writing sample have been normalized.
P-values are from the paired two-tailed t-test, with statistically significant results (at the 10% level) bolded.

Principles of economics classes often have many questions associated with mathematical computations and graphical solutions. Table 3 displays the performance of both cohorts on questions broken out by question type. The average on math questions with the writing cohort was slightly higher than the average of the non-writing cohort, but the difference was not statistically significant.

The fact that students were allowed to utilize calculators on each test may have mollified the effect of writing on the test, as students may directly enter the numbers for a problem into their calculator. This practice would reduce the effect writing on the test would have on problems with mathematical computations. We did find there was a statistically significant difference between the cohorts for questions that involved a figure or graph. We find that students who were able to write on the test performed significantly better on figure/graph question types compared to students in the non-writing cohort. This finding makes sense, as quite often figures and graphs are directly modified by the students in the process of deriving the solution to a problem. The evidence in Table 3 leads us to reject the figures and graphs portion of hypothesis 3 (students in each cohort (writing and no writing) will demonstrate no significant difference in test performance on questions categorized by the use of figures/graphs or mathematical computations) and fail to reject the math portion of hypothesis 3.

Table 3. Performance by Question Type.

Question Type	No Writing	Writing	Difference	p-values
Math (n=29)	71.9%	73.5%	+1.6%	0.3697
Figure/Graph (n=27)	64.2%	67.2%	**+3.1%**	**0.0152**

Note. All mean performances for the writing sample have been normalized.
P-values are from the paired two-tailed t-test, with statistically significant results (at the 10% level) bolded.

Finally, we examine whether allowing students to write on the test allows them to better answer questions that require higher-order thinking based on Bloom's taxonomy. Table 4 displays performance for the four levels of modified Bloom's taxonomy provided by the textbook: (1) Definitional, (2) Interpretive, (3) Applicative, and (4) Analytical. We find that there is a statistically significant improvement on question types that are definitional, interpretive, and applicative and we find no statistically significant effect on questions classified as analytical. The percentage of improvement decreases as the level of Bloom's taxonomy increases. This lends support to the idea that writing on the test helps eliminate mistakes on questions that require lower-level thinking rather than helping with questions that require deeper levels of thought. Thus, writing on the test helps eliminate the "careless mistakes" that students often lament making. The evidence in Table 4 leads us to reject hypothesis 4 (students in each cohort will demonstrate no significant difference in test performance on questions categorized by the publisher by the level of Bloom's taxonomy) for all levels of Bloom's categorization except analytical.

Table 4. Performance by Bloom's Taxonomy.

Bloom's Level	No Writing	Writing	Difference	p-values
Descriptive (n=32)	65.9%	72.1%	**+6.2%**	**0.0001**
Interpretive (n=44)	66.0%	70.6%	**+4.5%**	**0.0006**
Applicative (n=56)	64.1%	67.0%	**+3.0%**	**0.0017**
Analytical (n=18)	49.4%	47.1%	-2.2%	0.3139

Note. All mean performances for the writing sample have been normalized.
P-values are from the paired two-tailed t-test, with statistically significant results (at the 10% level) bolded.

Conclusions and Directions for Future Research

In this paper we propose that the testing mode effect should be broken into four distinct components: the effect that the ability to physically write on the test has on student performance, the effect of the method of answer recording has on student performance, the effect the proctoring/test-taking environment has on student performance, and the effect the test-taking environment has on the questions selected/used by the instructor for the test. We utilized an innovative normalization approach to test the first component of the testing mode effect: the effect that allowing (or disallowing) writing on the test has on student test performance. We found that students did statistically better when they were allowed to write on the test. In addition, we broke down student performance by question type (math, figure/graph, or neither), question difficulty (easy, moderate, and challenging), and level of Bloom's taxonomy (descriptive, interpretive, applicative, and analytical). Using a p-value of 0.10 or lower, we found that writing on the test statistically significantly enhances student performance in all areas except for questions classified as math, challenging, and analytical. Thus, we found broad support for the notion that permitting writing on the test allows students to perform at a higher level than when writing on the test is prohibited. It is important to note that these results hold despite the fact that students in the no-writing group were provided with scratch paper to write out whatever work they wish.

These findings are an important first step in ultimately deconstructing the effects of in-person paper-based tests that allow students to write on the test versus remotely proctored tests given on a computer. These results suggest that student performance in online and other non-writing environments may be degraded relative to in-person writing environments, and the instructor should be aware of (and perhaps compensate for) the adverse effect of this testing condition. In addition, the provision of scratch paper does not appear to be an effective substitute for allowing students to write on the test. This finding is especially important in an online setting, as many instructors allow scratch paper for online tests, perhaps falsely believing that it serves as a good substitute for paper testing. Given that online testing effectively prevents students from writing on the test, the instructor must be very careful when equating the results of online testing (or other no-writing testing) to paper-based testing. Perhaps future technological advances may bring about more effective countermeasures to this inherent limitation of online testing.

We also find that it is critically important to control for student aptitude across samples when studying the effect of various testing or teaching conditions. Given the relatively small samples involved with such studies, failure to account for underlying student differences can result in spurious conclusions being drawn. It is relatively difficult and data-intensive to control for students by using demographics in the traditional manner. In contrast, designing strong controls into the study from the beginning and utilizing a normalization procedure removes the need to impose complicated ex-post controls.

Future work should be devoted to further examining the effects of each of the four performance differentials of the testing effect. By deconstructing the effect of each of these changes individually, we can better understand how the testing mode affects student performance. With this knowledge, we may be able to craft tests and testing conditions that are more "mode neutral" in accordance with the tenets of good instructional design as laid out by Clark (1994) and the International Testing Commission (2006). This would help ensure that students in all modes of delivery would be treated equally and would have the best opportunity to demonstrate their true level of knowledge on a test. This is important if we wish to be consistent in our administration and assessment of testing material across different modes of delivery and assessment. Finally, additional work should be devoted to deconstructing the effect of the testing environment on students by demographic characteristics. This would allow us to identify if changes in the testing environment

disproportionately affect students from various demographic and socioeconomic backgrounds.

Acknowledgements

We thank the faculty and students in the College of Business and Industry at the University of Arkansas-Fort Smith who provided valuable comments during the presentation of preliminary results. In addition, we express our gratitude to the reviewers and editors who enhanced the final version of this paper.

References

Abdul, B., Adesope, O.O., Thiessen, D.B., & Wie, B.J.V. (2016). Comparing the effects of two active learning approaches. *International Journal of Engineering Education 32*(2), 654–669.

Backes, B. & 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. https://doi.org/10.1016/j.econedurev.2018.12.007

Bretz, S.L., & McClary, L. (2015). Students' Understandings of Acid Strength: How Meaningful Is Reliability When Measuring Alternative Conceptions? *Journal of Chemical Education 92*(2). 212–219. https://doi.org/10.1021/ed5005195

Clariana, R., & 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. https://doi.org/10.1111/1467-8535.00294

Clark, R.E. (1994). Media will never influence learning. *Educational Technology Research and Development 42*(2). 21–29. https://doi.org/10.1007/BF02299088

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics 66*(1). 64–74. https://doi.org/10.1119/1.18809

Hensley, K. K. (2015). Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement. The University of Iowa.

The International Test Commission. (2006). International Guidelines on Computer-Based and Internet-Delivered Testing. *International Journal of Testing 6*(2). 143-171. http://dx.doi.org/10.1207/s15327574ijt0602_4

Katz, I.R., Bennett, R.E., & Berger, A.E. (2000). Effects of Response Format on Difficulty of SAT-Mathematics Items: It's Not the Strategy. *Journal of Educational Measurement 37*(1). 39–57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Marshman, E., & Singh, C. (2016). Interactive tutorial to improve student understanding of single photon experiments involving a Mach–Zehnder interferometer. *European Journal of Physics 37*(2). 1-22. https://doi.org/10.1088/0143-0807/37/2/024001

Prisacari, A.A., & Danielson, J. (2017 a). Rethinking testing mode: Should I offer my next chemistry test on paper or computer? *Computers & Education 106*. 1–12. https://doi.org/10.1016/j.compedu.2016.11.008

Prisacari, A.A., & Danielson, J. (2017 b). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior 77*. 1–10. https://doi.org/10.1016/j.chb.2017.07.044

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the Comparability of Paper-and Computer-Based Science Tests Across Sex and SES Subgroups. *Educational Measurement: Issues and Practice 31*(4). 2–12.

Settlage, D.M., & Wollscheid, J.R. (2019). An Analysis of the Effect of Student Prepared Notecards on Exam Performance. *College Teaching 67*(1). 15–22. https://doi.org/10.1080/87567555.2018.1514485

Sherman, T.J., Harvey, T.M., Royse, E.A., Heim, A.B., Smith, C.F., Romano, A.B., King, A.E., Lyons, D.O., & Holt, E.A. (2019). Effect of quiz format on student performance and answer-changing behaviour on formative assessments. *Journal of Biological Education 55*(3): 1–15. https://doi.org/10.1080/00219266.2019.1687106

Siegfried, J.J., & Kennedy, P.E. (1995). Does Pedagogy Vary with Class Size in Introductory Economics? *The American Economic Review 85*(2). 347–351.

Smolinsky, L., Marx, B.D., Olafsson, G., & Ma, Y.A. (2020). Computer-Based and Paper-and-Pencil Tests: A Study in Calculus for STEM Majors. *Journal of Educational Computing Research 58*(7). 1256–1278. https://doi.org/10.1177/0735633120930235

Vispoel, W.P., Morris, C.A., & Clough, S.J. (2019). Interchangeability of Results From Computerized and Traditional Administration of the BIDR: Convenience Can Match Reality. *Journal of Personality Assessment 101*(3). 237–252. https://doi.org/10.1080/00223891.2017.1406361

Wainer, H., & Thissen, D.B. (1993). Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education 6*(2). 103–118. https://doi.org/10.1207/s15324818ame0602_1

Walstad, W.B. (1998). Multiple choice tests for the economics course. In W. B. Walstad & P. Saunders (Eds.), *Teaching undergraduate economics: A handbook for instructors*, 287–304. McGraw - Hill.

Walstad, W.B., & Becker, W.E. (1994). Achievement Differences on Multiple-Choice and Essay Tests in Economics. *The American Economic Review 84*(2). 193–196.