November 2016

Thank you for your request to our REL Reference Desk regarding **the effectiveness of online assessments vs paper assessments. Ask REL Southwest** is part of a collaborative Ask-A-REL reference desk service provided by the 10 regional educational laboratories (RELs). By design, this service functions much in the same way as a technical reference library, by providing references, referrals, and brief responses in the form of citations for research-based education questions.

Please note that REL Southwest has not done an evaluation of the resources themselves but offers this list to you for your information only.

BACKGROUND

State accountability testing in recent years has been moving from to paper-pencil to computer-based mode. A subsequent question is: Do the computer-based exams that are increasingly prevalent in K-12 education measure skills and knowledge as accurately as traditional paper-based tests? The difference between computer- and paper-based testing is often a matter of personal preference, personality, and familiarity with the format.

Following an established REL Southwest protocol, we conducted a search for research reports, websites, as well as descriptive briefs on the effectiveness of online assessments versus paper assessments. The sources included federally funded organizations, additional research institutions, educational databases, and general Internet searches using Google and Bing. See the methods section at the end of this Ask A REL for additional information on how we identified the following sources.

QUESTION

What does research tell us about the effectiveness of online assessments versus paper assessments?

SOURCES

Edwards, V. B. (2014). Digital advances reshaping k-12 testing. *Technology Counts*. *Education Week*. v 33, n 25. http://eric.ed.gov/?id=ED561009.

From the ERIC abstract: “Figuring out how to use digital tools to transform testing requires a willingness to invest in new technologies and the patience to experiment with novel approaches, a commitment to ongoing professional development and reliable technical support, and an openness to learn from mistakes. Whatever bumpy ride this technological journey takes, experts insist that online assessments--for both high-stakes tests and classroom exams--are the undeniable wave of the future. They see online

tests, and adaptive ones in particular, as a key tool for building personalized learning programs that address students' individual strengths and weaknesses. And with only about a year to go before students in most states are scheduled to take new, online assessments aligned to the Common Core State Standards districts are still taking stock of whether the technology they have on hand will meet their needs. Many schools are now seeing, late in the game, that the gap between what they have and what they need is troubling. A recent report by the State Educational Technology Directors Association, in Glen Burnie, MD., suggests that concerns about schools' technological readiness for common-core testing are justified. It found that 72 percent of schools do not meet the basic Internet-bandwidth requirements of 100 kilobits per second per student set by the association-- essentially the minimum of what's required for a schoolwide 1-to-1 computing environment. That lack of preparation for the common-core online tests could be a major missed opportunity for many schools, experts point out, because the digital upgrades put in place for the common core could fuel the use of technology to transform testing in other ways. Having a stronger technology backbone in place could also set the stage for wider use of assistive technologies. Once seen as primarily for students with disabilities those technologies are now merging into the broader testing world, especially as more states and districts embrace online testing. Computer-based exams provide an opportunity to allow all students to tap into accommodations that could aid comprehension and focus. The rich multimedia content and interactive experiences in games and simulations provide an opportunity for deeper insights into the nuances and complexities of how students solve problems. Even so, some experts advise schools to stay focused on integrating technology into assessments in thoughtful ways that have an impact on learning. "Technology Counts 2014: Digital Advances Reshaping K-12 Testing" examines technology developments that have prompted a rethinking of assessments. Discover what districts are doing to find the technology that fits their testing needs, particularly in relation to new common assessments. Articles in this issue include: (1) Testing Digital Advances (Kevin C. Bushweller); (2) Building Better Feedback Loops (Benjamin Herold); (3) "Playlists" Tailor Curriculum (Benjamin Herold); (4) Automating Writing Evaluations (Caralee J.Adams); (5) Assistive Tech for Everyone? (Michelle R. Davis); (6) New Tools Evolve to Address Autism (Michelle R. Davis); (7) Testing Students in Simulated Worlds (Benjamin Herold;(8) Moment of Truth for Common Core (Sean Cavanagh); (9) Collaborating on Testing (Sean Cavanagh); (10) Taking the Pulse of Digital Literacy (Robin L. Flanigan); (11) Districts Tackle Technology Gaps (Amanda M. Fairbanks); (12) Preventing Digital Cheating (Michelle R. Davis); and (13) Playing Games, Evaluating Skills (Robin l. Flanigan)."

Foorman, B., Espinosa, A., Wood, C. and Wu, Yi-C. (2016). Using computer-adaptive assessments of literacy to monitor the progress of English learner students. REL 2016-149, *Regional Educational Laboratory Southeast*. http://eric.ed.gov/?id=ED566912.

From the ERIC abstract: "A top education priority in the United States is to address the needs of one of the fastest growing yet lowest performing student populations--English learner students (Capps et al., 2005). English learner students come from homes where a non-English language is spoken and need additional academic support to access the

mainstream curriculum. These students account for about 10 percent of the preK-12 student population in the United States (Aud et al., 2013). Spanish-speaking students account for 80 percent of the English learner student population in the United States and, because they live disproportionately in poverty and attend schools with higher percentages of racial/ethnic minority students, students from low-income households, and students with low achievement, Spanish-speaking students are at greater risk of low achievement than other English learner students (Capps et al., 2005). This study examined how teachers and school staff administered computer-adaptive assessments of literacy to English learner students in grades 3-5 and how they used the assessments to monitor students' growth in literacy skills. It presents findings that may aid districts in implementing a computer-adaptive assessment of literacy skills for English learner students as well as for other students. Appendix A presents Additional Tables and Figures."

Gullen, K. (2014). Are Our Kids Ready for Computerized Tests? *Educational Leadership*, v71 n6 p68-71. http://eric.ed.gov/?id=EJ1043841.

From the description: "As standardized assessments tied to the Common Core standards approach for K-12 students, U.S. teachers correctly feel that how we test students will change. Will students be ready for tests of proficiency done on computers? Gullen debriefed with 500 students in various grades who had just taken pilot assessment items connected to the coming assessments online, asking them what was positive or difficult about doing the tests online and what skills they thought students would need to take such tests. Third graders had trouble with using a cursor and scroll bar, secondary students reported trouble negotiating large amounts of text on-screen, and all students were frustrated by having to type their answers--so that they often cut short their responses. Gullen recommends six practices teachers should take to help students strengthen both computer skills and the ability to work independently on lengthy problems with many parts."

Herold, B. (2016). Comparing Paper and Computer Testing: 7 Key Research Studies, Education Week, v 35, Issue 22, p 8. http://www.edweek.org/ew/articles/2016/02/23/comparing-paper-and-computer-testing-7-key.html .

From the article: "To give a deeper look at the issues behind this "mode effect," *Education Week* examined seven key research studies on the topic:

1. "Online Assessment and the Comparability of Score Meaning" **Educational Testing Service, 2003—**[http://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf].
"It should be a matter of indifference to the examinee whether the test is administered on computer or paper, or whether it is taken on a large-screen display or a small one," wrote Randy Elliot Bennett more than a decade ago. Bennet was one of the leaders in the field of psychometrics and mode-comparability, and this overview explores a range of mode-comparability issues. 'Although the promise of online assessment is substantial, states are encountering significant issues, including ones of measurement and fairness,' the paper reads. 'Particularly

distressing is the potential for such variation [in testing conditions] to unfairly affect population groups, such as females, minority-group members, or students attending schools in poor neighborhoods.'

2. "Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches, and Trends"
Pearson, 2008—
[http://images.pearsonassessments.com/images/tmrs/Maintaining_Score_Equivalence_as_Tests_Transition_Online.pdf].
The authors of this paper, originally presented at the National Council of Measurement in Education, highlight the "mixed findings" from studies about the impact of test-administration mode on student reading and mathematic scores, saying they "promote ambiguity" and make life difficult for policymakers. The answer, they say, is quasi-experimental designs carried out by testing entities such as state departments of education. The preferred technique, the paper suggests, is a matched-samples comparability analysis, through which researchers are able to create comparable groups of test-takers in each mode of administration, then compare how they performed.

3. "Does It Matter If I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP"
Journal of Technology, Learning, and Assessment, 2008—
[http://files.eric.ed.gov/fulltext/EJ838621.pdf].
"Results showed that the computer-based mathematics test was significantly harder statistically than the paper-based test," according to Randy Elliot Bennett, who is also the lead author of this paper, which looked at results from a 2001 National Center for Education Statistics investigation of new technology for administering the National Assessment of Educational Progress in math. 'In addition, computer facility predicted online mathematics test performance after controlling for performance on a paper-based mathematics test, suggesting that degree of familiarity with computers may matter when taking a computer-based mathematics test in NAEP.'

4. "The Nation's Report Card: Writing 2011"
National Center for Education Statistics, 2014—
[http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf].
As the NCES moved to administer its first computer-based NAEP writing assessment, it also tracked the impact in this study of how 24,100 8th graders and 28,1000 12th graders performed. Doug Levin, then the director of the State Educational Technology Directors Association, summed up the findings in a 2014 blog post: 'Students who had greater access to technology in and out of school, and had teachers that required its use for school assignments, used technology in more powerful ways' and 'scored significantly higher on the NAEP writing achievement test,' Levin wrote. 'Such clear and direct relationships are few and far between in education—and these findings raise many implications for states and districts as they shift to online assessment.'

5. "Performance of 4th-Grade Students in the 2012 NAEP Computer-Based Writing Pilot"
NCES, 2015—[http://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf].
This working paper found that high-performing 4th graders who took NAEP's computer-based pilot writing exam in 2012 scored "substantively higher on the computer" than similar students who had taken the exam on paper in 2010. Low- and middle-performing students did not similarly benefit from taking the exam on computers, raising concerns that computer-based exams might widen achievement gaps. Likely key to the score differences, said Sheida White, one of the report's authors, in an interview, is the role of "facilitative" computer skills such as keyboarding ability and word-processing skills. 'When a student [who has those skills] is generating an essay, their cognitive resources are focused on their word choices, their sentence structure, and how to make their sentences more interesting and varied—not trying to find letters on a keyboard, or the technical aspects of the computer,' White said.

6. "Mathematics Minnesota Comprehensive Assessment-Series III (MCA-III) Mode Comparability Study Report"
Minnesota Department of Education and Pearson, 2012—
[http://blogs.edweek.org/edweek/DigitalEducation/MathMCA-III_ModeComparabilityStudy.pdf].
This state-level study of mode effects on exams administered in spring and summer of 2011 used the matched-samples comparability-analysis technique described in the Pearson study. 'Although the results indicated the presence of relatively small overall mode effects that favored the paper administration, these effects were observed for a minority of items common to the paper and online forms; the study found.

7. "Comparability of Student Scores Obtained From Paper and Computer Administrations"
Oregon Department of Education, 2007—
[http://www.ode.state.or.us/teachlearn/testing/manuals/2007/doc4.1comparabilitytestatopandp.pdf]
This state-level mode-comparability study looked across math, reading, and science tests administered by both computer and paper. 'Results suggest that average scores and standard errors are quite similar across [computer] and paper tests. Although the difference were still quite small (less than a half a scale score point), 3rd graders tended to show slightly larger differences,' the paper reads. 'This study provides evidence that scores are comparable across [Oregon's computer] and paper delivery modes.'"

Stone, E. and Davey, T. (2011). Computer-Adaptive Testing for Students with Disabilities: A review of the literature. Research Report. ETS RR-11-32, *ETS Research Report Series*. http://eric.ed.gov/?id=EJ1110441.

From the ERIC abstract: "There has been an increased interest in developing computer-adaptive testing (CAT) and multistage assessments for K-12 accountability assessments. The move to adaptive testing has been met with some resistance by

those in the field of special education who express concern about routing of students with divergent profiles (e.g., some students with math-based learning disabilities may have difficulty with basic computation but not high level problem solving) and poor performance on early test questions. This paper consists of a literature review focusing on adaptive testing issues for students with disabilities in the K-12 sector. While it is clear that there are issues that will present obstacles to administering accountability tests adaptively to students with disabilities, this synthesis of research and policy developments with respect to this topic will be useful both for development of research agendas and to inform states that are currently using or are considering moving to CAT."

Thissen, D. and Norton, S. (2013). What might changes in psychometric approaches to statewide testing mean for NAEP? American Institutes for Research. http://eric.ed.gov/?id=ED545241.

From the ERIC abstract: "Development of the Common Core State Standards (CCSS), and the creation of the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC), changes the pattern of accountability testing. These changes raise the question: "How should NAEP's validity and utility be maintained?" The assessments planned by the consortia may be different enough from current state assessments to raise questions as to whether NAEP can continue to play its historic role as an independent monitor or "check" on the validity of state assessments. It is also clear is that computer-based assessment is coming to K-12 education, and both consortia plan to include more varied item types than have been commonly used in the past. Computerization of NAEP is inevitable and already planned by the National Assessment Governing Board. Computerized NAEP assessments may appear more similar to future statewide assessments. Comparability of results can usually be maintained as a test makes the transition from paper-and-pencil to computerized administration, but computerization may have an effect on results for some subgroups of the population. Computerization of NAEP is best approached in the same way as other changes to NAEP assessments have been approached: A bridge study should insure the comparability of results across the transition unless an a priori decision is made to "break trend" regardless. Assessments developed by Smarter Balanced and PARCC may reduce the number of statewide tests to the low single digits, thus making linkage feasible. Associations between the results of disparate educational assessments tend to change over time, so any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly. A singular opportunity exists in a short window of time--essentially right now--to design the data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development. Two appendices present: (1) Membership in the PARCC and Smarter Balanced Consortia; and (2) Computer-Based Assessment: A Review of the Last 15 Years of Comparability Research. [For the main report, "Examining the Content and Context of the Common Core State Standards: A First Look at Implications for the National Assessment of Educational Progress," see ED545237.]"

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, v67 n2 p219-238. http://eric.ed.gov/?id=EJ757687.

From the ERIC abstract: "This study conducted a meta-analysis of computer-based and paper-and-pencil administration mode effects on K-12 student mathematics tests. Both initial and final results based on fixed- and random-effects models are presented. The results based on the final selected studies with homogeneous effect sizes show that the administration mode had no statistically significant effect on K-12 student mathematics tests. Only the moderator variable of computer delivery algorithm contributed to predicting the effect size. The differences in scores between test modes were larger for linear tests than for adaptive tests. However, such variables as study design, grade level, sample size, type of test, computer delivery method, and computer practice did not lead to differences in student mathematics scores between computer-based and paper-and-pencil modes. (Contains 5 tables.)"

Woods, Julie Rowland (2015). Testing trends: Considerations for choosing and using assessments, Education Commission of the States. http://eric.ed.gov/?id=ED561807.

From the ERIC abstract: "While federal law requires students to be tested in math, English-language arts and science in particular grades, states are still struggling to mount the resources and expertise necessary to fully implement college and career readiness standards, let alone new assessments aligned to these higher standards. New assessments are not only more demanding of students but also come with new administrative, technological and scoring challenges. Policymakers, caught between testing and accountability requirements and their constituents' concerns, are seeking new ways to meet the needs of all stakeholders. The key questions faced by education leaders are: (1) Which assessments should we choose? and (2) How do we use those assessments? To aid education leaders and policymakers in answering these questions, this report highlights how other states have addressed these questions and their attendant issues." NOTE: This source was not peer-reviewed."

ADDITIONAL ORGANIZATIONS AND RESOURCES TO CONSULT

Burndette II, D. (2016). Online-Testing Stumbles Spark Legislation in Affected States, *Education Week*. http://www.edweek.org/ew/articles/2016/03/09/online-testing-stumbles-spark-legislation-in-affected.html.

This article discusses the issues many states are having with online state accountability tests.

From the article: "The sometimes-temporary, but ill-timed glitches have fueled legislation that would crack down on testing companies and have lasting impact on the role tests play in evaluating schools and teachers."

Cavanaugh, Cathy; Gillan, Kathy Jo; Kromrey, Jeff; Hess, Melinda; Blomeyer, Robert (2004). The effects of distance education on K-12 student outcomes: A meta-

analysis, *Learning Point Associates / North Central Regional Educational Laboratory (NCREL)* http://eric.ed.gov/?id=ED489533.

From the ERIC abstract: "The community of K-12 education has seen explosive growth over the last decade in distance learning programs, defined as learning experiences in which students and instructors are separated by space and/or time. While elementary and secondary students have learned through the use of electronic distance learning systems since the 1930s, the development of online distance learning schools is a relatively new phenomenon. Online virtual schools may be ideally suited to meet the needs of stakeholders calling for school choice, high school reform, and workforce preparation in 21st century skills. The growth in the numbers of students learning online and the importance of online learning as a solution to educational challenges has increased the need to study more closely the factors that affect student learning in virtual schooling environments. This meta-analysis is a statistical review of 116 effect sizes from 14 web-delivered K-12 distance education programs studied between 1999 and 2004. The analysis shows that distance education can have the same effect on measures of student academic achievement when compared to traditional instruction. The study-weighted mean effect size across all outcomes was -0.028 with a 95 percent confidence interval from 0.060 to -0.116, indicating no significant difference in performance between students who participated in online programs and those who were taught in face-to-face classrooms. No factors were found to be related to significant positive or negative effects. The factors that were tested included academic content area, grade level of the students, role of the distance learning program, role of the instructor, length of the program, type of school, frequency of the distance learning experience, pacing of instruction, timing of instruction, instructor preparation and experience in distance education, and the setting of the students. Appended is: Coded Variables and Study Features in the Codebook."

Herold, B. (2016). PARCC Scores Lower for Students Who Took Exams on Computers, *Education Week,* v 35, Issue 20, pp 1, 11. http://www.edweek.org/ew/articles/2016/02/03/parcc-scores-lower-on-computer.html.

From the article: "Students who took the 2014-15 PARCC exams via computer tended to score lower than those who took the exams with paper and pencil—a revelation that prompts questions about the validity of the test results and poses potentially big problems for state and district leaders."

Karkee, T., Kim, D. and Fatica, K. (2010). Comparability study of online and paper and pencil tests using modified internally and externally matched criteria. Paper presented at the annual meeting of the American Educational Research Association (AERA), Denver, CO April 29 – May 4, 2010. http://www.measurementinc.com/sites/default/files/Online%20and%20Paper%20and%20Pencil%20Comparability%20Study%20with%20Alternate%20Design.pdf.

From the abstract: "Many states are exploring possibilities of transitioning to online mode of test administration with the premise that future tests will be computer based of some sort. The benefits include quick on-demand reporting, immediate feedback, and

cost effectiveness. Comparability studies are conducted to help demonstrate and ensure the defensibility of using the scores interchangeably between paper and pencil and online tests. Common designs used in comparability study require either random assignment of examinees or assign testing mode to random equivalent groups.

Alternate designs conditioned on internal and internal plus external criteria for creating equivalent samples are compared in this study together with the scores between modes of administration with in a condition. The mode effects at item and student level were evaluated by comparing model fit, differential item functioning, and mean of item and person parameters. The test results did not show statistically discernible mode effects based on model fit, DIF, or student performance, despite some differences in item parameters."

Kaufman, J., Hamilton, L., Stecher, B., Naftel, S., Robbins, M. Garber, C., Ogletree, C., Faxon-Mills**,** S., and Opfer, V. D. (2016). What are teachers' and school leaders' major concerns about new K–12 state tests? Findings from the American **t**eacher and American school leader panels. Santa Monica, CA: *RAND Corporation*. http://www.rand.org/pubs/research_reports/RR1294-1.html.

Among other findings, this report stated that more teachers perceived that their schools' technological capacity would not be sufficient to administer the English Language Arts (ELA) and math tests. The percentage of teachers reporting concerns about Partnership for Assessment of Readiness for College and Careers (PARCC) was significantly higher ($p<0.05$) than for teachers reporting on other assessments.

From the report description: "Many states have recently made major changes to their K–12 student testing programs. The media have reported growing dissatisfaction with the amount of testing happening in schools and the use of tests for high-stakes decisionmaking. However, there is little systematically gathered information on the perspectives of U.S. educators who have firsthand knowledge about testing and its effects on teaching and learning. This report shares U.S. principals' and teachers' main concerns about testing, drawing upon new survey tools for understanding educators' perspectives and implementation of major education policies: RAND's American Teacher Panel (ATP) and American School Leader Panel (ASLP). The findings are drawn from the ATP and ASLP surveys fielded in February 2015, before the full administration of most state-mandated exams. Findings indicate particular concern with students' test performance, as well as more prevalent concerns about the PARCC assessment compared with other assessments. The information about U.S. educators' concerns will serve as a baseline for tracking changes in attitudes over time. This analysis focuses on "the main state-mandated test for mathematics" and "for English language arts" (ELA) that teachers and principals reported their students taking. This report was updated in October 2016. The current version provides estimates based on updated weights for a small percentage of the respondents. Weights were updated to account for infrequent misclassification in the assignment of school-level characteristics.

Key Findings

- Most teachers expressed moderate or major concerns about test difficulty, low student performance on the tests, and test score accuracy for special needs students.
- Teachers with students taking Partnership for Assessment of Readiness for College and Careers (PARCC) tests were more likely to be concerned about testing issues than teachers with students taking other state tests.
- Teachers at low-income schools were more likely to be concerned about testing issues than teachers in other schools."

Olson, L. (2005). Impact of Paper-and-Pencil, Online Testing Is Compared, *Education Week* http://www.edweek.org/ew/articles/2005/08/31/01online.h25.html.

From the article: "How students perform on computer-delivered tests depends, in part, on how familiar they are with the technology, concludes a set of studies conducted by the Princeton, N.J.-based Educational Testing Service.

The studies looked at how students performed when given mathematics and writing items from the National Assessment of Educational Progress by paper and pencil vs. computer."

Computer vs Paper by Enoch Morrison—
https://www.testprepreview.com/computer_paper.htm—Last Updated: 09/30/2016

METHODS

Keywords and Search Strings Used in the Searches:

Paper vs computer tests; computer-assisted testing; student evaluation; Online standardized assessment; K-12 online assessment; K-12 student performance on online assessments OR tests

Search of Databases and Websites

- Institute of Education Sciences (IES) website (http://www.ies.ed.gov) and IES sources: Regional Educational Laboratory (REL) Program, National Center for Education Statistics (NCES), National Center for Education Research (NCER), What Works Clearinghouse (WWC)
- ERIC database (www.eric.ed.gov)
- Google Scholar (scholar.google.com)
- Google (www.google.com)
- Bing (www.bing.com)

Criteria for Inclusion

REL Southwest selected resources that provide research on the effectiveness of online assessments vs paper assessments. When REL Southwest staff reviewed resources, we considered – among other things – three factors:

1. **Date of Publication:** The most current information (primarily published from 2010 to the present) is included.
2. **Source and Funder of the Report/Brief/Article:** Priority was given to publications written in relevant, peer-reviewed journals or reports or produced by well-known research organizations.
3. **Methodology:** Sources include reported studies, literature reviews and policy reports.

Ask A REL is a service provided by a collaborative of the Regional Educational Laboratory (REL) Program, funded by the U.S. Department of Education's Institute of Education Sciences (IES). This response was prepared by REL Southwest under contract ED-IES-12-C-0012 with IES. The content of this document does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.