

**Evaluation of the First Year Online
Michigan Educational Assessment Program Administration**



Office of Educational Assessment and Accountability
Michigan Department of Education
PO Box 3008
Lansing Michigan, 48909
www.michigan.gov/oeaa

June 8, 2006

Executive Summary

There are many benefits claimed for the use of online testing. This evaluation will assess the impact and efficacy of a pilot program on the use of online testing for the Michigan's assessment program. The major findings were:

- Analyses of student scores indicated that the online and paper versions of each test were sufficiently comparable.
- Assigning online constructed response scores based on the higher of automated and human ratings reduced the performance differences so that the students participating in the pilot were not disadvantaged.
- Comparisons between automated and human scores of the online constructed response items indicated that the agreement of the automated and human ratings can be increased significantly by improving the representativeness of the constructed response responses used to train the automated scoring engine
- Overall, student reactions to online testing were overwhelmingly positive.
- Overall, faculty reactions were also overwhelmingly positive.
- Survey results and performance on the constructed responses indicated that some students were not comfortable with writing essays online.

The Online MEAP Pilot Program

The purpose of this evaluation was to determine whether providing the online MEAP test via online is:

- 1) at minimum equivalent to the same test and scoring results as the paper and pencil version or a consistent score that could be converted to an equivalent score,
- 2) determine if computer scored essays are at least equivalent to human scoring,
- 3) examine the technological capabilities of our schools to determine if schools have the readiness for this type of testing,
- 4) determine if there are advantages to this approach such as turnaround time on scores,
- 5) determine if students are capable of taking the exam using this method and how they perceive this type of testing, and
- 6) determine what kinds of administrative procedures are necessary at the state and local level to prepare for test administration and if the use of computers saves any efficiency in administrative procedures and staff time and resources.

This pilot online testing was conducted for the grade 6 English Language Arts and Social Studies tests administered under conditions (including test window) that were as similar as possible for the traditional MEAP program. The Freedom to Learn (FTL) program volunteered and paid for this study through a grant from Ferris State University College of Education (fiduciary). The study drew from schools that have been provided with laptop computers as part of the Freedom to Learn initiative. Leslie Wilson, Director, Professional and Curriculum Development of the Freedom to Learn Program sponsored by Ferris State University assisted MDE and PEM in identifying appropriate schools.

The online tests in English Language Arts and Social Studies were administered to grade 6 students using the wireless laptop computers that have been assigned to them. FTL was ideally suited for this project because of the leadership provided by FTL, and the partnerships between the school administration, teachers and local technical support.

Evaluation Design

There were 2,403 online tests administered to students in 19 Michigan school districts using Pearson Educational Measurement's (PEM) TestNav internet assessment system. Participating schools tested online during the same administration window (October 3 to October 21) as the paper-and-pencil MEAP.

Since this was an initial pilot and there were many unknown variables, it was decided to not introduce testing accommodations into this study. Provisional (raw) scores on the online tests were provided within 48 hours of the online administration. These included multiple-choice total scores and scores on grade-level expectations. Preliminary scores on constructed response (CR) items for the online tests were generated using Pearson Knowledge Technologies' (PKT)

automated essay scoring engine, referred to as the Intelligent Essay Assessor (IEA <http://www.knowledge-technologies.com>). However, CR items were also scored along with responses based on the paper test administration using traditional essay scoring procedures prior to final score reporting.

Test proctors were told: students should be familiar with the computer they will be using and should have keyboard skills sufficient to enter a one page answer to an essay question. The program includes a simple editing system that is easy to use. Students will not have to use a more complicated word processor and will not have to worry about formatting beyond paragraphs. In keeping with MEAP administration rules, there will be no spelling or grammar checking built-in to the software nor allowed to be used during the test.

An equating process was employed by PEM on the performance of an equivalent matched group of students from the regular MEAP test administration (matched for grade level and key demographic characteristics from their prior 5th Grade Social Studies Test performance) using exactly the same set of test items. This is a commonly used design for studying whether scores based on online and paper administrations of the same test are comparable.

To score essays automatically, PKT entered handwritten responses from field testing using a word processor so that those responses can be used to calibrate (or “train” the automated scoring engines. PKT utilized random samples of at least 200 field test responses and the scores assigned to these responses to calibrate the scoring engine. This engine was to be used to score the essays entered as part of the online pilot.

Surveys of students were taken online at the end of the assessment. Faculty and staff also received a survey several months later after they had a chance to review and use the preliminary results. A video based focus group sponsored by MASA was also completed. The faculty survey is included in Appendix A.

An extensive comparability study was conducted by Dr. Denny Wey of PEM. Some of the results will be reported here. The full report is included in Appendix B.

Findings

The primary results of the study are highlighted below:

- Analyses of student scores indicated that the online and paper versions of each test were sufficiently comparable for the same score conversion tables to be used in both modes.
- Analyses of the CR items in reading (one six-point item) and writing (one six-point item and one four-point item) indicated lower performance for students testing online versus students testing by paper. Assigning online CR scores based on the higher of automated and human ratings reduced the

performance differences so that the students participating in the pilot were not disadvantaged.

- Comparisons between automated and human scores of the online CR items indicated that the agreement of the automated and human ratings can be increased significantly by improving the representativeness of the CR responses used to train the automated scoring engine. In a follow up study where the automated scoring engine was trained using a representative sample of the operational online essays, agreement rates with human scores increased by 7 percent to 25 percent across the four CR items.
- Overall, student reactions to online testing were overwhelmingly positive. Over 90 percent of the online students agreed or strongly agreed with the statement, “Overall, I am happy that I took this test on computer rather than on paper” and over 90 percent indicated that they would recommend or strongly recommend taking the test by computer to other students.
- Overall, faculty reactions were very similar to the students. About 93% of faculty respondents either highly recommended or recommended with some reservations, continuing online testing. Only one faculty member did not recommend continuing it.
- Survey results indicated that some students were not comfortable with writing essays online. Only 30 percent of ELA respondents agreed or strongly agreed with the statement, “It is easier for me to write an essay on paper than on the computer.” When asked, “How often do you use a computer for writing papers or essays?”, 26 percent of ELA students indicated either “less than once a month” or “never.”
- There were some technical difficulties, though expected in this first attempt at online testing. While all schools managed to accomplish online testing, two schools had issues related to slow loading times for the test questions.
- The slow loading was attributed to two potential sources; 1) it was unclear to what extent the wireless configurations and router were responsible vs. 2) the secondary configuration of placing the software on the server (to facilitate a very last minute software update) may have contributed to slowing down the question download speed.

The overall raw scores of the tests were comparable when using the decision rule to use the higher score of the IEA vs. human scoring (see Table 1 in the PEM evaluation report). The scores were so comparable, that no adjustments for raw to scaled score conversions were needed.

The human scoring method and the IEA were not in the desired range of inter-rater agreement. An analysis by Dr. Wey (included in Appendix B (pg.20)) concludes that for this particular scoring rubric that the sample of pilot field test results (n=150) were insufficient to train the IEA engine. Increasing the field test samples sizes to 200 and 500 yielded inter-rater agreement level to at least equal or even higher levels of agreement when compared to human scoring.

Faculty and students were both extremely positive in their recommendations for using this assessment method. In both student surveys and in the faculty survey, over 90% of respondents stated they were positive or inclined to recommend online assessment.

Technology and Implementation Issues. There were two issues of concern in the use of technology; 1) keyboarding skills in relation to answering the constructed response, and 2) slow response of TestNav on some test sites.

One area of concern that was suspected to have a negative impact on a student's score on constructed response questions was poor keyboarding skills. While the demands of TestNav do not require touch typing, nor any type of formatting or word processing skills, they do require some reasonable speed. This was defined as being able to locate letters in a reasonable amount of time and knowing the functions of the backspace and enter/return key and simple cutting and pasting. This was expressed by the fact that in the student survey 30% of students responding strongly agreed or agreed with the statement, "It is easier for me to write an essay on paper than on computer." This was corroborated by the focus groups with school staff and by the open-ended questions from the faculty online survey.

The second area of concern that was brought up by schools in the survey and the focus groups was the slow response of TestNav in a few school buildings. This seemed to be an isolated problem, only a few schools made mention of this and did not appear to be a predominant issue. We were not able to ascertain if these were the largest schools in the sample. One potential solution may have inadvertently caused the slower response rate. Because of a need to upgrade the software to accommodate our project, the software had to be upgraded the weekend before the tests were to be offered live. Therefore the recommendation by PEM was to have one copy of the software reside on the server and place shortcuts on each computer used for testing. This allowed district technology staff to make one quick upgrade on the morning of the first day of testing. It may also have had the effect of slowing down the network in accessing student testing software loading during test administration. While this is our most likely theory; (one district did indicate that test loading speed had improved after loaded TestNav on each computer), we did not have an adequate amount of data on the problem to be sure that this was a complete explanation.

Discussion and Recommendations

Using the Freedom to Learn program turned out to be a very good way to introduce online high stakes testing to Michigan. The leadership from the program offices and the teamwork between the school administration and instructional staff and the technology staff should be used as a role model for other schools that wish to have a successful implementation of online testing.

Overall the biggest issue remains a concern about online constructed response and scoring. It was found that the small sample of field test scores were inadequate

for training the engine and when 200 and 500 samples produced IEA scores had equal to or higher inter-rater agreement than paper and pencil. Since this was empirically determined, the question of whether this will be true under actual online assessment operational conditions still remains.

One remaining area of concern is for students who did not appear to have the keyboard skills (roughly 30% by student report) to do the constructed response comfortably). This was compounded by a report in the faculty survey and focus group that some students did not know they could scroll down to add additional information to their answers. This suggests two possible problems; students did not have enough practice with CR before they took the test, and directions regarding how to make a complete answer (scrolling) need to be made clear.

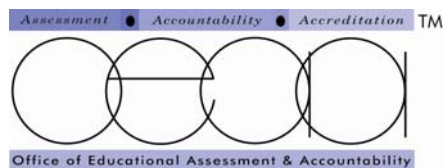
In summary,

- 1) The online and paper and pencil tests appeared to have comparable results.
- 2) The IEA computer scoring engine was trained with an inadequate sample.
 - i. The students scores in the pilot were not compromised because the higher of the two scores taken yielded equivalent results, and
 - ii. Using 200 and 500 samples to train the engine and then scoring led to equal or better comparability to hand scoring methods. This will not be a problem in the future since the embedded field tests will have significantly higher field test samples than 500.
- 3) The schools in this study appeared to have the technological capacity to conduct testing, however the ability to assess students in large numbers via wireless online testing using TestNav software loaded locally remains an open question.
- 4) Many administrators and teachers found the 48 hour turnaround on instructional scores useful, and there was significant decrease in administrative resources needed when compared to paper and pencil.
- 5) Students were mostly capable of taking the test online and were overwhelmingly positive about it. Students with poor keyboard skills continue to be an area of concern for online constructed response questions.
- 6) Appropriate and early training and the use of an e-manual combined with practice tests that include a CR question is necessary to improve online test administration practices.

As a result of this evaluation, the following is recommended:

- 1) A small task force advisory committee should be put together to consider whether online assessment is a long-term viable strategy and if so, how to proceed with a five year rollout plan.
- 2) There is a need for very clear directions for e-test delivery. A more detailed technical manual is necessary.
- 3) Technical support has to work very closely with the test administration staff in all phases of this project.

- i. The use of TestNav needs to be tested in advance of student examination delivery to determine if appropriate test loading speeds are available.
- 4) Students should have some level of keyboard skills prior to taking the constructed response questions. Those students who are not comfortable with minimal keyboarding (defined as the ability to type 2 or 3 paragraphs in a reasonable amount of time without having to search for each letter and knowledge of the backspace/delete key and the enter/return key as well as cut and paste) should take the test by paper and pencil.
- 5) Students need a good practice exam prior to MEAP administration that includes a CR with test directions that emphasize knowing how to scroll below the screen fold.



By
Paul M. Stemmer, Jr., Ph.D. and Joseph Martineau, Ph.D.
Office of Educational Assessment and Accountability
Michigan Department of Education
PO Box 30008
Lansing, MI 48909
www.michigan.gov/oeaa

Appendix A.

See MEAP Final Comparability Report Final 07/27/06