April 5, 2016

Too Many Unanswered Questions about Machine Scoring of the Common Core Exams

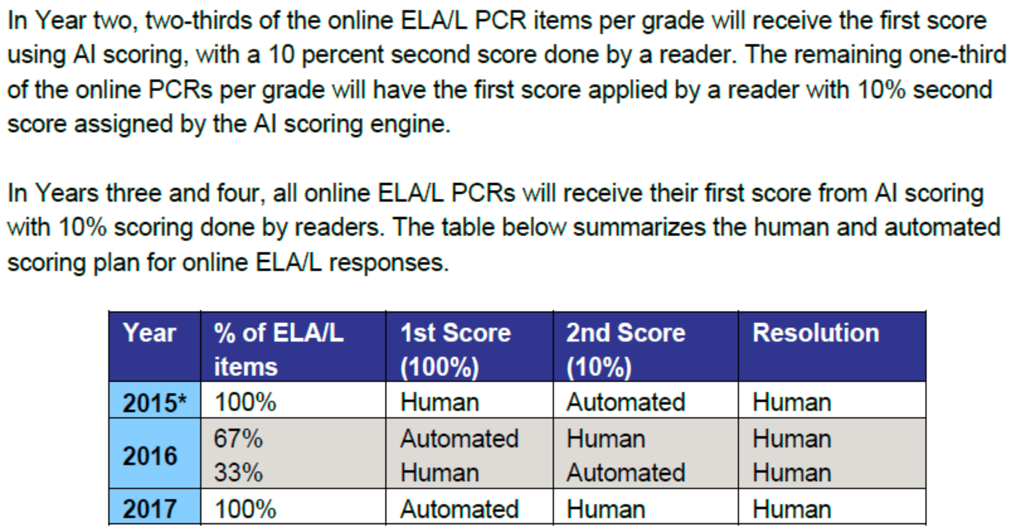
By Leonie Haimson, Executive Director, Class Size Matters and Co-chair, Parent Coalition for Student Privacy

The Common Core PARCC and SBAC exams have begun in 26 states across the country. According to Politico, two thirds of students' written responses in "many" PARCC states will be scored by computers, with only ten percent of those responses then checked by hand. That means a student has a little more than 36 percent chance of having their essays read by an actual human being. As we shall see, only 50 percent of students who take the SBAC exam are going to have their writing read and scored by humans.

After a sharp drop-off, PARCC states currently include Colorado, District of Columbia, Illinois, Maryland, New Jersey, New Mexico, and Rhode Island. (In Louisiana and Massachusetts, the Bureau of Indian Education and Department of Defense schools also are also participating to varying extents.)

Which poses the question, in which of these PARCC states will the vast majority of ELA tests be scored by machines? And where is the evidence that automated scoring is either valid or reliable? I and my colleagues in the Parent Coalition for Student Privacy did a little digging and we found out a little more, but not enough to answer either of these critical questions.

We discovered the following passages from the PARCC Ohio contract (although Ohio has now pulled out of PARCC because of huge technological glitches) and the Colorado contract isstill in force:



In Year two, two-thirds of the online ELA/L PCR items per grade will receive the first score using AI scoring, with a 10 percent second score done by a reader. The remaining one-third of the online PCRs per grade will have the first score applied by a reader with 10% second score assigned by the AI scoring engine.

In Years three and four, all online ELA/L PCRs will receive their first score from AI scoring with 10% scoring done by readers. The table below summarizes the human and automated scoring plan for online ELA/L responses.

Year	% of ELA/L items	1st Score (100%)	2nd Score (10%)	Resolution
2015*	100%	Human	Automated	Human
2016	67% 33%	Automated Human	Human Automated	Human Human
2017	100%	Automated	Human	Human

We also found this passage in the Executive Summary written by Pearson, the main contractor to PARCC:

> *Handscoring costs make up nearly 50 percent of the costs of the PARCC assessments. We are eager to conduct the automated scoring efficacy study later this year, in coordination with ETS*

> *and PARCC, with the plan to phase in automated scoring beginning with the spring administration of the first operational year.*

In another section of the contract, this "efficacy" or "Proof-of-concept" study on automated scoring to be carried out by Pearson mentioned was mentioned, with specific deadlines attached:

> *31. Automated Scoring: PARCC acknowledges the potential advantages of automated scoring ("AI Scoring") to promote efficiency associated with scoring of student constructed responses, which otherwise require human scoring. The automated scoring phase-in plan is incorporated in the base contract, as detailed in Table G-1 above and in the Redlined Proposal. However, Contractor shall conduct an efficacy study, and PARCC shall review contractors Proof-of-concept study results, and provide Contractor with authorization to utilize AI Scoring as specified herein. Contractor shall provide the following deliverables, as specified in Section V.B.3 of the Redlined Proposal:*
>
> *a. Proof-of concent research design*
>
> > *i. Report of results of proof-of-concept study, which shall be provided pursuant to the following Key Milestones and requirements: Proof-of-concept research design approved (to be mutually determined prior to baseline of project schedule.*
> > *ii. Contractor will provide proof-of-concept/efficacy study report to PAR; 10/15/14*
> > *iii. PARCC provides final approval to proceeded [sic] ith automated scoring phase-in plan based on results of efficacy study: 10/31/14*
> > *iv. Any states electing to opt out of phase-in plan and use human scoring for ELA online responses, notified Contractor no later than 11/14/14; signed agreement must be in place by 12/31/14*
> > *v. Any modification of the Key Milestone due dates, or change to the phase in plan, shall require a Scope Change pursuant to the terms of the Contract.*

So where is this proof-of-concept/efficacy study?

According to a Politico story from November 2014, though Pearson was supposed to deliver the study by mid-October 2014, the company missed that deadline.

As described above, the delivery of the study was supposed to trigger the final approval of the PARCC consortium to go ahead with automated scoring by October 31, 2014. If the deadline was missed this was supposed to require a "Scope Change" to the terms.

What happened when Pearson didn't deliver the contract? According to Politico,

> *There have been no scope changes to this portion of the contract, according to Larry Behrens, a spokesman for the New Mexico Public Education Department, which oversees the contract. So where's the Pearson study? PARCC spokesman David Connerty-Marin told Morning Education it's*

> *being revised — but he declined to say who had asked for the revisions or what they entail. Pearson wouldn't answer any questions on the subject, referring them all to PARCC.*
>
> *— **And what of the provision that PARCC provide "final approval"** of the phase-in of automated grading? Connerty-Marin wouldn't answer questions about whether a vote has already taken place or will be held in the future. He at first told Morning Education that PARCC states are currently "conducting studies" on the efficacy of using computer algorithms. But he later acknowledged that states aren't doing their own studies; they're relying on the Pearson report. Morning Education contacted all the states using PARCC tests to ask if they had made a decision on automated grading. Only Colorado and D.C. replied; both said no decision had been made.*

There is an alternative listed in the PARCC contracts, called the "Human Scoring Option" which would cost an additional $3.50 per student this year, rising to more than $5 per student in years three and four – nearly a 10% increase in price*.*

On March 2, 2015, PARCC put out a bulletin encouraging states to participate in their spring field testing "*to obtain item level data to support test construction for 2015–16 and gather responses to train the automated KT [apparently Pearson's Knowledge Analysis Technologies ™ (**KAT**)] scoring engine*. "

So just a little more than a year ago, Pearson was still "training" its system through the PARCC field testing program – which does not lend confidence to its proven efficacy or validity.

So was this "proof of concept" or efficacy study ever delivered to the states since the Politico article was published in November 2014? And if so, what did it say? On the PARCC website, there is a list of reports and studies that were completed, including this one:

> ***Automated Scoring Proof of Concept Study.*** *The purpose of the study was to evaluate whether machine scoring can be used in scoring of Prose Constructed Response (PCR) in PARCC ELA/Literacy assessments.*

We filled out an online form to ask for the study, but have not yet received a response.

An recent piece by Pearson on Artificial Intelligence (or AI) on this issue concludes, "*If we are ultimately successful, AIEd will also contribute a proportionate response to the most significant social challenge that AI has already brought – the steady replacement of jobs and occupations with clever algorithms and robots."*

What about the scoring method used by the Smarter Balanced exam, which is being given in 17 states this year, including California, Connecticut, Delaware, Hawaii, Idaho, Iowa, Montana, Nevada, New Hampshire, North Dakota, Oregon, South Dakota, Vermont, Washington, West Virginia and Wyoming, plus the US Virgin Islands, as well as to some extent in Michigan.

In an article dated March 15, 2013, Smarter Balanced reported that the consortium had retreated from its original plans to switch entirely to automated scoring:

> "Smarter Balanced has *actually already scaled back its plans for grading writing with machines because artificial intelligence technology has not developed as quickly as it had once hoped. In*

> *2010, when it was starting to develop the new Common Core exams for its 24 member states, the group wanted to use machines to grade 100 percent of the writing.*
>
> *"Our initial estimates were assuming we could do everything by machine, but we've changed that," said Jacqueline King, a director at Smarter Balanced. The technology hasn't moved ahead as fast as we thought," King said.*

Yet here is an excerpt from the Connecticut agreement with AIR, the primary contractor of the Smarter Balanced exam:

AGREEMENT BETWEEN
THE CONNECTICUT STATE BOARD OF EDUCATION AND THE AMERICAN INSTITUTES
FOR RESEARCH

i. During Year 1, AIR shall hand-score 100 percent of the responses and do human second double-blind scorings 15 percent of the time. AIR shall use Artificial Intelligence (AI) to conduct a second score of the remaining 85 percent of the responses.

ii. In Year 2, AIR shall use AI to score 100 percent of the responses and shall hand-score 50 percent on a second scoring.

iii. For Year 3 and beyond, AIR shall use AI to score 100 percent of the responses shall hand-score25 [sic] percent on a second scoring.

So given that it is now year two, only half of the students in Connecticut will have the chance for a human to read their responses.

This AIR agreement also references a validity study that was supposed to have been provided to states concerning the PEG AI [Artificial Intelligence] scoring system to be used by SBAC: "*a report containing multiple measures of human/AI agreement, including: percent perfect agreement, percent adjacent, percent perfect + adjacent, and quadratic weighted kappa*".

According to experts, the purpose of the second human scoring is to confirm the accuracy of machine scoring. Without making public the degree to which both are correlated, there is no evidence that the machines are able to come close to replicating human scores.

The PEG AI scoring system elsewhere is described as "an automated scoring technology … purchased by Measurement Incorporated in 2002." On another PEG site, it says: "*Although factors like creativity are beyond the scope of computerized assessment, these programs can still be used in the classroom to provide timely, unbiased feedback that allows students to improve their writing skills and proficiency*."

Which brings up the obvious point – wasn't the Common Core supposed to encourage creativity and critical thinking? And the Common Core aligned exams supposed to assess these skills? Is there any evidence that machines can do either? As far as one can determine, the answer is no.

Last year, Les Perelman, who was in charge of MIT's Writing program, wrote an opinion piece for the Boston Globe. Perelman tested out another automated scoring system, IntelliMetric, that could not distinguish essays with meaningful coherent prose from nonsense, and that high marks to gibberish, such as this:

> *"According to professor of theory of knowledge Leon Trotsky, privacy is the most fundamental report of humankind. Radiation on advocates to an orator transmits gamma rays of parsimony to implode.''*

Unable to analyze meaning, narrative, or argument, automated scoring instead relies on length, grammar, and measures of abstruse vocabulary to do assess prose.

Perelman had asked to test the Pearson *KAT* system being used by PARCC, but was denied access to their robo-grader. He concluded:

> *If PARCC does not insist that Pearson allow researchers access to its robo-grader and release all raw numerical data on the scoring, then Massachusetts should withdraw from the consortium. No pharmaceutical company is allowed to conduct medical tests in secret or deny legitimate investigators access. The FDA and independent investigators are always involved. Indeed, even toasters have more oversight than high stakes educational tests…*

A paper dated March 2013 from the Educational Testing Service (one of the SBAC sub-contractors) concluded:

> *Current automated essay-scoring systems cannot directly assess some of the more cognitively demanding aspects of writing proficiency, such as audience awareness, argumentation, critical thinking, and creativity. …A related weakness of automated scoring is that these systems could potentially be manipulated by test takers seeking an unfair advantage. Examinees may, for example, use complicated words, use formulaic but logically incoherent language, or artificially increase the length of the essay to try and improve their scores.*

According to a recent article, Florida plans to use AIR's AutoScore to grade essays on its statewide exams; and Utah has reportedly used automated scoring since 2010. As of March 2015, New Jersey had not yet made up its mind whether to adopt automated scoring for the PARCC exams to be given this spring.

Last year, the NJ Department of Educator director of assessments Jeff Hauger was quoted as saying the state had not decided how quickly to adopt computer scoring:

> *"The state will consider the option if the automated scoring proves to be accurate and cost effective. … But the state understands the perception that automated scoring may not be as effective as being graded by hand…We would not go full automated scoring without having some information for us to believe that actually it does just as good of a job as human scores."*

Given all the unresolved issues, today parent leaders and advocates from many states, including Parent Coalition for Student Privacy, Parents Across America, FairTest and Network for Public Education, sent a letter to their State Education Chiefs, demanding answers to the following questions:

1- What percentage of the ELA exams in our state are being scored by machines this year, and how many of these exams will then be re-scored by a human being?
2- What happens if the machine score varies significantly from the score given by the human being?
3- Will parents have the opportunity to learn whether their children's ELA exam was scored by a human being or a machine?
4- Will you provide the "proof of concept" or efficacy studies promised months ago by Pearson in the case of PARCC, and AIR in the case of SBAC, and cited in the contracts as attesting to the validity and reliability of the machine-scoring method being used?
5- Will you provide any independent research that provides evidence of the reliability of this method, and preferably studies published in peer-reviewed journals?

We encourage parents to send their own letters to their State Education chiefs, as well as state and local school officials, asking these questions. And if you don't get answers to your questions, or answers you think are sufficient, you should definitely consider opting your child out of these exams.

There are also significant issues about the lack of privacy afforded your child's personal data by PARCC, SBAC and their numerous contractors and subcontractors. PARCC has an extremely weak privacy policy. As far as we have been able to determine, SBAC has no privacy policy at all –at least one they have made available to parents.

More on automated scoring: A ***petition*** *posted by academics against automated scoring of high stakes exams has collected* ***4304 signatures*** *since March 12, 2013 including Noam Chomsky (who knows a little about language); urging policymakers and assessment designers including PARCC and SBAC to* "stop *using invalid computerized scoring of student essays." It also cites research findings and bibliography on the issue.* *Here's a scholarly critique by Les Perelman of the claims made by proponents of machine scoring, who cite a 2012 competition sponsored by the Hewlett Foundation. Here's a NY Times column about the controversy.*

Other readings:

Bennett, Randy Elliot. "The changing nature of educational assessment." *Review of Research in Education* 39.1 (2015): 370-407.

Perelman, L. "Construct Validity, Length, Score, And Time In Holistically Graded Writing Assessments: The Case Against Automated Essay Scoring." *International Advances in Writing Research: Cultures, Places, Measures.* Ed. Charles Bazerman, Chris Dean, Jessica Early, Karen Lunsford, Suzie Null, Paul Rogers, and Amanda Stansell. Colorado: The WAC Clearinghouse and Parlor Press, 2012. 121-132.

Perelman, L. "Mass-Market Writing Assessments as Bullshit." *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White*. Ed. Norbert Elliot and Les Perelman. Hampton Press, 2012. 425-437.

Perelman, L. "When 'the state of the art' is counting words." *Assessing Writing* 21 (2014): 104-111.