Assessing Elementary Students' Science Competency with Text Analytics

Samuel P. Leeman-Munk
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695
spleeman@ncsu.edu

Eric N. Wiebe
Department of STEM Education
North Carolina State University
Raleigh, North Carolina 27695
wiebe@ncsu.edu

James C. Lester
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695
lester@ncsu.edu

ABSTRACT

Real-time formative assessment of student learning has become the subject of increasing attention. Students' textual responses to short answer questions offer a rich source of data for formative assessment. However, automatically analyzing textual constructed responses poses significant computational challenges, and the difficulty of generating accurate assessments is exacerbated by the disfluencies that occur prominently in elementary students' writing. With robust text analytics, there is the potential to accurately analyze students' text responses and predict students' future success. In this paper, we present WRITEEVAL, a hybrid text analytics method for analyzing student-composed text written in response to constructed response questions. Based on a model integrating a text similarity technique with a semantic analysis technique, WRITEEVAL performs well on responses written by fourth graders in response to short-text science questions. Further, it was found that WRITEEVAL's assessments correlate with summative analyses of student performance.

Categories and Subject Descriptors

K.3.m [**Computers and Education**]: Miscellaneous.

General Terms

Human Factors

Keywords

Text-based learning analytics, formative assessment, automated assessment, constructed response analysis, writing assessment

1. INTRODUCTION

Recent years have seen a growing interest in investigating how student learning data can be analyzed in real-time for automated formative assessment to support teachers in the classroom [11, 17]. A broad base of research in science education and other STEM fields has been investigating the role of formative assessment in instruction [1, 6, 11]. This work makes clear that the more restrictive methods traditionally used in summative assessment such as multiple choice questions are limited in their ability to provide the analyses necessary for guiding real-time scaffolding and remediation for students (e.g., [3]). To address this issue, recent approaches to real-time formative assessment have included analyses of student action logs in an open-ended learning environment [15] and analyses of interactions with course materials and online tools to predict student performance [2, 20].

As a tool for formative assessment, short-text constructed response items reveal cognitive processes and states in students that are difficult to uncover in multiple-choice equivalents [13]. Even when it seems that items could be designed to address the same cognitive construct, success in devising multiple-choice and constructed-response items that behave with psychometric equivalence has proven to be limited [9]. Because standards-based STEM education in the United States explicitly promotes the development of writing skills for which constructed response items are ideally suited [12, 14, 16], the prospect of designing text analytics techniques for automatically assessing students' textual responses has become even more appealing, leading various groups to pursue research in the area [5, 8, 10].

An important family of short answer questions is the constructed response question. A *constructed response question* is designed to elicit a response of no more than a few sentences and features a relatively clear distinction between incorrect, partially correct, and correct answers. Ideally, a system designed for *constructed response analysis* (CRA) would be machine-learned from examples that include both graded student answers and expert-constructed "reference" answers [4]. The challenges of creating an accurate machine-learning-based CRA system stem from the variety of ways in which a student can express a given concept. In addition to lexical and syntactic variety, students often compose ill-formed text replete with ungrammatical phrasings and misspellings, which significantly complicate analysis.

In this paper we present WRITEEVAL, a hybrid approach to constructed response analysis for student science responses. We also investigate whether WRITEEVAL's analyses of a student's work as she progresses through a problem-solving session can be used to predict her performance on a summative multiple-choice post-test. WRITEEVAL uses a hybrid model of two techniques: a text similarity technique (*soft cardinality*) and a semantic analysis technique (*precedent feature collection*). The precedent feature collection (PFC) technique learns directly from scores assigned by human graders and accounts for lexical variety using semantic comparison methods. Because neither technique relies on word order, they are robust to grammatical disfluencies. WRITEEVAL has been evaluated on a dataset of textual responses to short-text science questions collected in a study conducted at elementary schools in two states. Responders were in fourth grade and generally aged between nine and ten. The results indicate that

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA
Copyright 2014 ACM 978-1-4503-2664-3/14/03 $15.00.
http://dx.doi.org/10.1145/2567574.2567620

Table 1. Student Answers and Their Human-generated Grades According to the Science Score Rubric

Question	How can you make the two bulbs in a series circuit brighter? Write an answer to the focus question based on what you have already learned about simple and series circuits.	
Reference Answer	Add 2 D-cells to make the bulbs brighter.	**Grade**
Student Answer 1	Use a parrelel circiut its were two circiuts connect in a diffrent way	*Correct*
Student Answer 2	By having more then one battire then atching the wires to each light bolb	*Correct*
Student Answer 3	You can change it to a parallel circuit	*Correct*
Student Answer 4	Make the wire shorter.	*Partially Correct*
Student Answer 5	A new d cell	*Partially Correct*
Student Answer 6	Lite and a dcell	*Partially Correct*
Student Answer 7	Connect it in a easier way	*Incorrect*
Student Answer 8	I still think the same thing	*Incorrect*
Student Answer 9	By putting a wire conecting from a motar wire to a light bolb wire	*Incorrect*

even in the presence of high disfluency, WRITEEVAL shows promise for enabling a teacher to predict a student's future success with similar content.

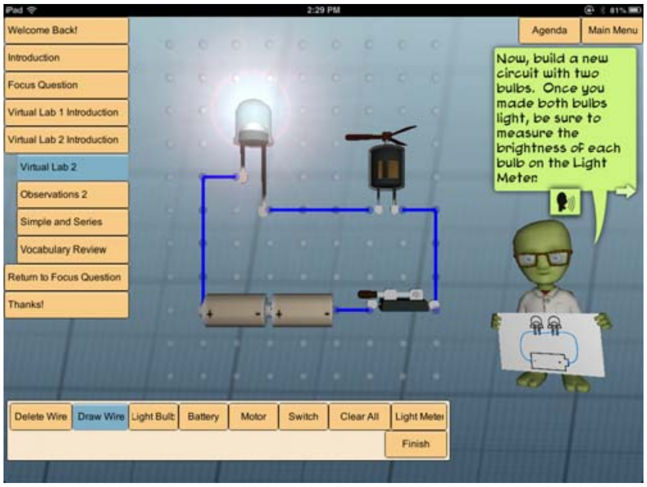
This paper is structured as follows. Section 2 introduces the tablet-based digital science notebook software that was used to collect the constructed response dataset from elementary students. Section 3 introduces WRITEEVAL and its underlying hybrid machine learning approach. Section 4 describes an evaluation of WRITEEVAL and investigates its ability to predict the learning trajectories of students. Section 5 discusses the findings and explores how WRITEEVAL may serve as the basis for future approaches to real-time formative assessment.

2. CONSTRUCTED RESPONSE DATA SET

Constructed response questions can play a central role in science assessment. We have been exploring constructed response assessment in the context of science education for upper elementary students with the LEONARDO CyberPad (Figure 1). Under development in our laboratory for three years, LEONARDO is a digital science notebook that runs on tablet computing platforms. LEONARDO integrates intelligent tutoring systems technologies into a digital science notebook that enables students to graphically model science phenomena. With a focus on the physical and earth sciences, the LEONARDO PadMate, a pedagogical agent, supports students' learning with real-time problem-solving advice. LEONARDO's curriculum is based on that of the Full Option Science System [19]. As students progress through the curriculum, they utilize LEONARDO's virtual notebook, complete virtual labs, and write responses to constructed response questions. To date, LEONARDO has been implemented in over 40 classrooms around the United States.

The short answer and post-test data used in this investigation were gathered from fourth grade students during implementations of LEONARDO in public schools in California and North Carolina. The data collection for each class took place over five days with students completing a new Energy and Circuits investigation each day. Two human graders graded students' responses from the dataset on a science score rubric with four categories: *no answer*, *incorrect*, *partially correct*, and *correct*. The graders graded one class of data and then conferred on where their results had disagreed. They then graded other classes. On a sample of 10% of the responses of the classes they graded after conferring, the graders achieved a Cohen's Kappa of 0.72.

Automatically grading students' responses is challenging for two reasons. First, students use phrasing and concepts that are difficult to anticipate in reference answers, and, second responses exhibit considerable disfluency. Table 1 displays a representative question, reference answer, and student responses. Note that Student Answer 1 features both correct answers unanticipated by the reference and significant disfluency. Student Answer 4 was deemed to be partially correct because although wires, which have non-zero resistance, do dim the light bulbs slightly when they are longer, the effect is unlikely to be noticeable. Even conventional answers, though, can be hard to detect due to disfluency such as misspellings. An analysis of constructed responses collected in this study reveals that 4.7% of words in all of student answers combined are not found in a dictionary. This is in contrast to other similar datasets, such as the Beetle dataset of undergraduate text answers to science questions, which features a 0.8% rate of out-of-dictionary words [4]. In each case, the numbers underestimate overall spelling errors. Misspellings such as 'batter' for 'battery', are not counted as missing in a dictionary test. These *real-word spelling errors* nevertheless misrepresent a student's meaning and complicate analysis. We describe how WRITEEVAL addresses these issues in Section 3.



Now, build a new circuit with two bulbs. Once you made both bulbs light, be sure to measure the brightness of each bulb on the Light Meter.

Figure 1. The LEONARDO CyberPad Digital Science Notebook

3. CONSTRUCTED RESPONSE ANALYSIS

WriteEval uses a hybrid machine learning approach that combines two complementary techniques to machine-learn how to grade student answers based on a corpus of human-graded answers. The two techniques are motivated by a need to be robust to unanticipated student answers and to disfluency. WriteEval's first technique, soft cardinality [7], uses decompositions of words into character sequences to identify the similarities between misspellings of the same word. Considering "dcells" in an example answer, "mor dcells," and "D-cells" in the reference answer, we can find overlaps in "ce," "el," "ll," "ls," "ell," "lls," and so on up to and including "cells." This technique functions equally well for real-word spelling errors such as if the student had forgotten the "d" and typed only "cells." Such overlaps signify a close match for both of these words. Soft cardinality generates features comparing the question and reference answer, the student answer and reference answer, and the student answer and the question.

WriteEval's second technique, Precedent Feature Collection (PFC), enacts machine-learned semantic comparisons to enable it to infer a correct answer not represented in the reference answer. Most notably, PFC can account for correct answers that are not present as reference answers, such as a "use a parallel circuit" for how to make two bulbs brighter when the reference answer is "add two batteries." PFC's semantic similarity measures are based on Latent Semantic Analysis (LSA). Trained on a corpus of relevant texts, LSA notes the co-occurrence of words and develops topics that describe sets of words that tend to occur together. In this manner, an answer can match another answer even if they share few or no words so long as their words represent similar topics. "D-Cell" and "Battery," for example, would likely appear in the same topic.

To mediate the need for a curriculum author to anticipate all possible correct answers in reference answers, PFC collects similarity features directly from a training set of graded student responses. PFC separates documents into groups by grade and takes each group as the "precedent" for that grade. Thus, Student Answers 1, 2, and 3 in Table 1 would all be in one group, as would Student Answers 4, 5, and 6 and 7,8, and 9.

To grade "parallel circuit," PFC compares the students' response with all five of the pre-graded answers and collects statistics on the similarity scores for each group. Because one of the answers in the Grade 3 precedent contains "parallel" and "circuit," which both match directly with the student answer, some of the similarity statistics in the Grade 3 precedent would reflect this, e.g., the mean and maximum similarities. As it is currently implemented, WriteEval ignores the ordering inherent in a numeric grading rubric like Science Score.When generating a grade, WriteEval is given a training set of pre-graded answers, the relevant question and reference answers, and the student's response to be graded. The student answer and the reference answer are passed to both soft cardinality and PFC; the question is passed to soft cardinality; and the human-graded answers are passed to PFC. Soft cardinality and PFC each return their own set of features, which are concatenated to form the set of features for the classifier.

WriteEval integrates the two components of its techniques as shown in Figure 2. In addition to the elements depicted in Figure 2, WriteEval also computes a similarity measure between the student answer and the reference answer, as well as the mean and sum grades weighted by similarity of the k most similar pre-graded answers. In addition to the precedents based on all pre-graded answers, WriteEval generates another set of precedents from answers with the same question as the answer to be graded.

To machine-learn its classifiers, WriteEval collects all the features generated by these techniques and feeds them into a decision tree algorithm. A decision tree algorithm tries to find conditions such as "is the similarity score to correct answers greater than 0.7," that maximally distinguish between scores. To classify an unseen answer, the classifier can follow the conditions until after a series of conditions it reaches the score that met all those conditions in the training set. To reduce the variability of our model we use 15-bagging, meaning our result comes from a vote between 15 trees each trained on a different random selection of the data. Our system uses the Weka J48-Graft decision tree algorithm, whose details are outside the scope of this paper [18].

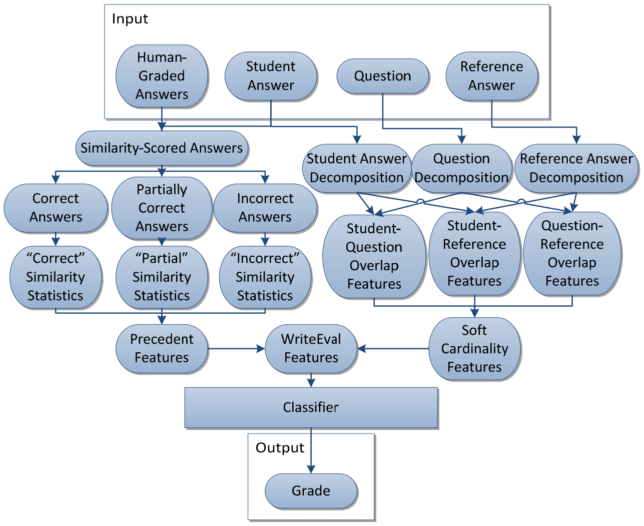

Input
Human-Graded Answers
Student Answer
Question
Reference Answer
Similarity-Scored Answers
Student Answer Decomposition
Question Decomposition
Reference Answer Decomposition
Correct Answers
Partially Correct Answers
Incorrect Answers
Student-Question Overlap Features
Student-Reference Overlap Features
Question-Reference Overlap Features
"Correct" Similarity Statistics
"Partial" Similarity Statistics
"Incorrect" Similarity Statistics
Precedent Features
WriteEval Features
Soft Cardinality Features
Classifier
Output
Grade

Figure 2. WriteEval Constructed Response Analysis

4. EVALUATION

WriteEval's performance was evaluated with 10-fold cross-validation using the dataset described above. With an accuracy of 68%, and a macro-averaged precision and recall of 68% and 57%, WriteEval significantly outperforms ($p < .001$) the majority class baseline, which would assign each answer in the dataset the most common score (*Partially Correct*), achieving an accuracy of 58.4% and a macro-averaged precision and recall of 19.4% and 33.3%. To determine the usefulness of automatic grading of science content in predicting the overall trajectory of a student's performance, we computed a running average of science scores on students' answers as they moved through the five-day Energy and Circuits lesson. We calculated the correlation between our running average of formative assessments and the student's score on a ten-point multiple-choice test taken after the five days of using Leonardo.

A critical assumption underlying our running average is that students answered each question in order. Although Leonardo does not prevent students from answering questions out of order, it is organized to strongly encourage linear progression. We excluded empty responses from the running average because it can be difficult to determine if a response was left empty because a student skipped a question, ran out of time, or was simply absent that day. Students who had responded to fewer than five out of twenty responses were dropped, and one question was a duplicate and dropped, leaving responses from sixty-seven students to at least five of nineteen questions each.

Figure 3 shows the correlation between the running average of automatic scoring by WriteEval and post-test scores, as well as that of expert human scoring. Starting with the response to the second question, the running average of automatic science scores correlates significantly ($p < .05$) with students' post-test scores. The correlation starts at .32 for two answers, but as we collect more observations, climbs up to .54. As it grades more questions, the running average of automatic science scores' correlation to students' post-test scores begins to converge with that of the running average of expert human scoring. WriteEval starts out about 0.2 lower than human judgments in correlation with students' post-test scores, but at Questions 9 and 12 our automatic scoring technique improves considerably. For more detail on the contributions of individual questions, Figure 4 compares their correlations with post-test score.

Figure 4 shows the ten questions with significant correlation to post-test scores, including Questions 9 and 12. Seven of these ten questions maintain their significant correlation when graded by WriteEval. One question, Question 19, saw a significant correlation between WriteEval-scored responses and post-test scores and not between human scored responses and post-test scores. This question was excluded from the chart because the significance was only very slightly below .05 (.045), and because it had the fewest student responses of any question, only twenty out of sixty-seven students answered question 19. Questions 9 and 12 are associated with considerable jumps in Figure 3, most likely due to their positioning. They successfully identified something about a student that will affect the student's eventual post-test performance and that prior questions had not yet assessed.

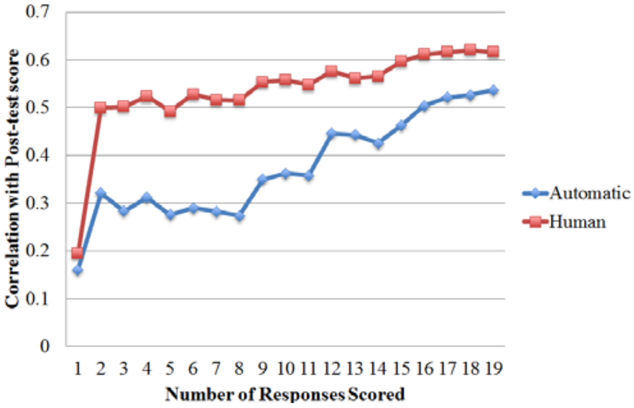


Figure 3: Correlation of the Running Average of WriteEval and Human-Graded Science Scores with Post-test Score

5. DISCUSSION

The evaluation suggests that WriteEval shows promise for generating real-time predictions of overall student performance. The significance of the correlation over the running average for all but the first question indicates the predictive potential of WriteEval, as does the tendency of the automatic and human correlations towards convergence as more questions are scored. WriteEval likely performed especially well on Questions 9, 10, and 12 because the key difference between a correct and incorrect answer was the presence or absence of certain key terms (e.g., 'nail' in Question 9 and 'metal' in Question 12). Although WriteEval performed especially well on these two questions, its overall performance indicates that it appears to be able to work effectively on many different kinds of questions.

Question 1's lack of correlation with post-test performance for both the human and automatic scores is likely related to the relative lack of range in student response quality. Nearly all students with responses to Question 1 received a score of either *Partially Correct* or *Correct* from the human graders. On the question, "What do you need to make a light bulb light?" it is easy for students to be assigned a grade of *Partially Correct* by mentioning any of the components of a circuit, which are extensively discussed in the preceding text, leading to this question being unreflective of students' overall ability regardless of the means of scoring it.

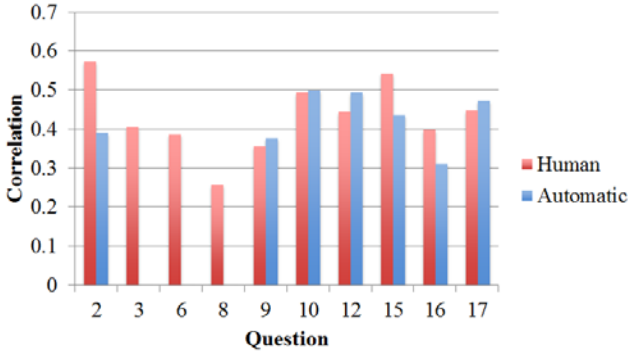


Figure 4: Correlation of Individual Questions with Post-test Score

By the time students have progressed halfway through the lesson (around Question 9), we can see a distinct pattern emerging. As we add more questions, the automatic score's correlation approaches closer and closer to the human scores' correlation with post-test score, which rises very slowly but steadily. If in a future study we were to extend to more questions, trends suggest we could expect to see asymptotic convergence between human and automatic scoring as well as a tapering off of the slow rise in predictiveness of human scores.

In these yield curves, we would anticipate that eventually the human scores would reach a stability point where additional questions would reveal effectively nothing more about a student's eventual success or failure. This would be the point, if not before, at which formative assessment by either the Leonardo PadMate or a teacher no longer yield new insight as to a student's overall understanding of the topic area. This yield curve would also need to take into consideration the practical limits of instructional time that could be devoted to CRA type work by the student. Field testing in classrooms would be necessary to fully realize the proper deployment of such an analytic tool so that a teacher is able to fully realize the diagnostic potential of this information.

6. CONCLUSION

This paper presents a hybrid text analytics method to support real-time formative assessment. Integrating two complementary methods, soft cardinality and a semantic analysis technique, the text analytics method has been implemented in WriteEval, a constructed response assessment system for text-based responses. WriteEval has been evaluated on highly disfluent constructed response texts composed by fourth grade students interacting with a tablet-based digital science notebook. The results of the evaluation suggest that WriteEval's hybrid machine-learned model generates assessments that are predictive of students' post-test performance. It offers the potential to produce assessments in real-time that may serve as early warning indicators to help teachers strategize as to how to allocate instructional interventions to support student learning.

WriteEval’s current performance levels suggest several promising directions for future work. First, it will be important to extend WriteEval’s ability to deal with responses exhibiting greater disfluency and including a greater number of unanticipated elements, particularly unanticipated elements that are also disfluent. Second, WriteEval should be extended to consider answers in more detail than simple assessment of correctness. Argumentation phenomena, which are particularly important in science education, will be a focus of future studies. Third, it will be instructive to incorporate WriteEval into the Leonardo digital science notebook to investigate techniques for classroom-based formative assessment that artfully utilize both intelligent support by the PadMate and personalized support by the teacher.

Reliable analysis of constructed response items not only provides additional summative analysis of writing ability in science, but also gives the teacher a powerful formative assessment tool that can be used to guide instructional strategies at either the individual student or whole class level. Given that time for science instruction is limited at the elementary level, the use of real-time assessment to address student misconceptions or missing knowledge immediately can be a valuable classroom tool.

7. ACKNOWLEDGMENTS

The authors wish to thank our colleagues on the Leonardo team for their contributions to the design, development, and classroom implementations of Leonardo: Courtney Behrle, Mike Carter, Bradford Mott, Angela Shelton, Peter Andrew Smith, and Robert Taylor. This material is based upon work supported by the National Science Foundation under Grant No. DRL1020229. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Abell, S. and Lederman, N. 2007. *Handbook of Research on Science Education*. Routledge. New York, NY.

[2] Arnold, K.E. and Pistilli, M.D. 2012. Course signals at Purdue. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12* (New York, New York, USA, 2012), 267–270.

[3] Bell, B. and Cowie, B. 2001. The characteristics of formative assessment in science education. *Science Education*. 85, 5 (2001), 536–553.

[4] Dzikovska, M., Nielsen, R. and Brew, C. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Montreal, Canada, 2012), 200–210.

[5] Graesser, A. 2000. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*. 8, 2 (2000), 1–33.

[6] Gunnarsson, B.L. and Alterman, R. 2012. Predicting failure: A case study in co-blogging. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12* (New York, New York, USA, 2012), 263–266.

[7] Jimenez, S., Becerra, C. and Gelbukh, A. 2013. SOFTCARDINALITY: hierarchical text overlap for student response analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (2013), 280–284.

[8] Jordan, S. and Butcher, P. Does the Sun orbit the Earth ? Challenges in using short free-text computer-marked questions .

[9] Kuechler, W. and Simkin, M. 2010. Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*. 8, 1 (2010), 55–73.

[10] Labeke, V. and John, T.E. 2013. OpenEssayist : extractive summarisation and formative assessment of free-text essays Conference Item. (2013).

[11] Monroy, C., Rangel, V.S. and Whitaker, R. 2013. STEMscopes: Contextualizing learning analytics in a K-12 science curriculum. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13* (New York, New York, USA, 2013), 210–219.

[12] NGSS Lead States 2013. *Next Generation Science Standards: For States, By States.* National Academic Press. Washington DC.

[13] Nicol, D. 2007. E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education*. 31, 1 (2007), 53–64.

[14] Porter, A., McMaken, J., Hwang, J. and Yang, R. 2011. Common core standards the new US intended curriculum. *Educational Researcher*. 40, 3 (2011), 103–116.

[15] Sao Pedro, M.A., Baker, R.S.J.D. and Gobert, J.D. 2013. What different kinds of stratification can reveal about the generalizability of data-mined skill assessment models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13* (New York, New York, USA, 2013), 190–194.

[16] Southavilay, V., Yacef, K., Reimann, P. and Calvo, R.A. 2013. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13* (New York, New York, USA, 2013), 38–47.

[17] Tempelaar, D.T., Heck, A., Cuypers, H., van der Kooij, H. and van de Vrie, E. 2013. Formative assessment and learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13* (New York, New York, USA, 2013), 205–209.

[18] Webb, G. 1999. Decision tree grafting from the all-tests-but-one partition. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (San Francisco, CA, 1999), 702–707.

[19] Welcome to FossWeb: 2013. *http://www.fossweb.com/*. Accessed: 2013-10-20.

[20] Wolff, A., Zdrahal, Z., Nikolov, A. and Pantucek, M. 2013. Improving retention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK ’13* (New York, New York, USA, 2013), 145–149.