Industry Track - Short Papers

Analysing and Refining Pilot Training

Bruno Emond
National Research Council Canada
1200 Montreal Road, Ottawa,
ON, Canada. K1A 0R6
1-613-993-0154
bruno.emond@nrc-cnrc.gc.ca

Scott Buffett
National Research Council Canada
46 Dineen Drive, Fredericton,
NB, Canada. E3B 9W4
1-506-444-0386
scott.buffett@nrc-cnrc.gc.ca

Cyril Goutte
National Research Council Canada
1200 Montreal Road, Ottawa,
ON, Canada. K1A 0R6
1-613-993-0805
cyril.goutte@nrc-cnrc.gc.ca

Ruibiao Jaff Guo
CAE Defense & Security
1135 Innovation Dr, Kanata,
ON, Canada. K2K 3G7
1-613-247-0342
jaff.guo@cae.com

ABSTRACT
Competency based training has become a major thrust in the development of instruction in both civilian and military pilot training. This paper reports on a joint effort by CAE and the National Research Council to identify data analytics methods relevant for the analysis, and refinements of competency based pilot training. In particular, these methods aim to identify correlations between 1) student actions and behaviours while engaging in training, and 2) students' success and incremental progression in the corresponding competencies being acquired. The paper presents some of our main results in applying sequence mining and additive factor modelling to small sets of pilot training data.

Keywords
Aviation pilots, competency-based training, sequence mining, additive factor models.

1. INTRODUCTION
Over the years, CAE has developed many research collaborations with universities and government research laboratories. The current paper presents some results from a project between CAE[1], the Advanced Technologies for Learning in Authentic Settings (ATLAS) research team from McGill University, and the Learning and Performance System Support program at the National Research Council Canada. The research efforts were focused on the identification of education data mining methods with practical outcomes for the improvement of pilot training. The main objective is to be able to analyse performance, and use competency models in order to refine simulation scenarios and CBT courseware. The contributions to the project represent different perspectives from sequence mining (descriptive method), to logistic regression models (predictive method). The objective was to explore the data from different points of view.

The following section presents an overview of the main trends in pilot training including competency, evidence, and scenario-based training. The next section briefly presents the data set that was used for all the analysis, and the remaining two sections presents the main results of applying sequence mining and additive factor modeling to this data.

[1] http://www.cae.com/about-cae/corporate-information/faq/

2. TRENDS IN PILOT TRAINING
To address the challenges of pilot training in the early 2000s, civil aviation stakeholders like the Civil Aviation Safety Alert (CASA), the International Civil Aviation Organization (ICAO), and concurrently the United States Air Force (USAF) have been promoting competency and evidence based training as a training model [1]–[3]. This position was in reaction to hours-based training where the number of flight hours or sorties done by a pilot determined flight or mission readiness. With the increase of flight operation complexities, it became obvious that achievement of a certain performance level on a task would be a better indication of a pilot competency, than the number of hours of practice, even though flight hours could be an indirect measure of a competency level.

There are many views about what a competency is. The International Civil Aviation Organization defines a competency as "a combination of skills, knowledge and attitudes required to perform a task to the prescribed standard" [4]. The USAF has developed an elaborate competency framework [5]. The Mission Essential Competencies (MEC) framework is intended to blend training task lists, and mission essential task lists. The MECs incorporate a wide range of pilot competencies, beyond the operational requirements, to include teams and inter-team competencies [3]. The Federal Aviation Administration (FAA) also recognizes that pilot competencies need to be defined at a higher-level than simply the low-level operations of an aircraft, especially with the increased level of automation because automated systems are not adapted to unforeseen situations [6]. Competency frameworks are usually the result of an analysis performed by subject matter experts who identify key competencies based on standards of performance and means to measure them.

Another important trend in pilot training is evidence-based training. The ICAO defines evidence-based training as "Training and assessment based on operational data that is characterized by developing and assessing the overall capability of a trainee across a range of core competencies rather than by measuring the performance in individual events or manoeuvres" [1]. The essential element evidence-based training introduces to competency based-training is the reference to operational data as a means to identify key competencies, in addition to the analysis

performed by subject matter experts. Evidence-based training applies the principles of competency-based training for safe, effective and efficient airline operations, while addressing safety threats. The term evidence refers to the fact that safety threats are identified from actual flight monitoring data, such as those provided by the Flight Operational Quality Assurance (FOQA) program, Aviation Safety Action Program (ASAP) data for business aviation [7], as well as Automatic Dependent Surveillance-Broadcast (ADS-B) data.

A literature review also revealed that a combination of competency, evidence, and scenario-based training approaches can form the basis for the next generation of pilot training system. The combination requires links between the development of simulated scenario events and performance measures, both driven by training objectives [8]**.** This combination is well integrated in the specification of evidence-based training as defined by the ICAO [1], and the focus on scenarios and simulations provides the foundation of a strong learner centred approach.

Simulation scenarios are central to evidence-based training as the main instructional content a trainee pilot interacts with, for evaluation and learning. The approach is consistent with the principles of situated learning theory, which argues that learning best takes place in the context in which it is going to be used. Scenario-based training is mostly suitable for procedure-oriented tasks requiring decision-making and critical thinking in complex situations, and is learner centered as the scenario provides a unique opportunity for the trainee to perform and acquire competencies based on his/her competency level.

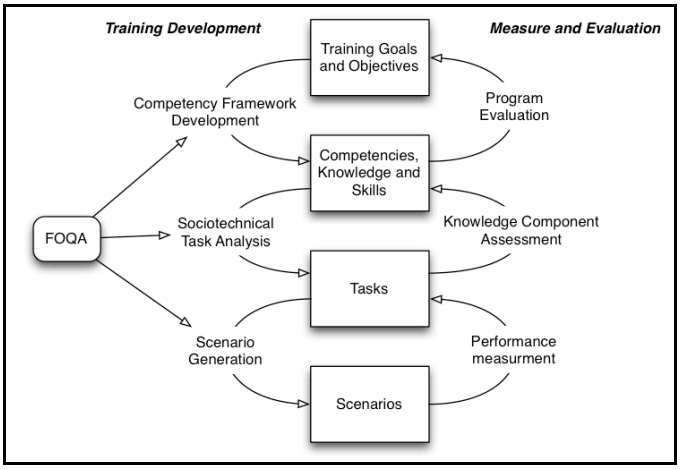


Figure 1. Competency, evidence and scenario-based training systems

Figure 1, inspired from [8]**,** tries to capture the relationships between competency-based training, evidence-based training as flight data monitoring programs feed in information for training development at all levels, and scenario-based training which constitutes an essential element for providing learner centered experiences. In addition to the closed workflow between A) training goals and objectives; B) competencies, knowledge, and skills; C) tasks; and D) scenarios, Figure 1 distinguishes on the left hand side training development including: the specification of competency frameworks, sociotechnical task analysis, and scenario generation. The right hand side of the figure presents key elements related to the measure and evaluation including: performance measurement, knowledge component assessment, and program evaluation.

The remaining sections of the paper fall essentially within the right hand side of Figure 1 under "Knowledge Component Assessment". The courseware delivery software gathered the student learning performance data during the learning process, including the sequences of activities selected by the students, timestamps, and question answers.

3. DATA DESCRIPTION

The data consists of two sets of web training sessions engaging students on scenarios requiring information gathering, review and assessment of new flight procedures with demands on both knowledge and skill acquisition related to taking off and landing operations. The two data sets correspond to two separate groups of students, and had respectively eight and six students in them. Table 1 presents the frequency distribution of events either as being assessments or information-gathering events for each student in the two groups. The counts in Table 1 refer to the sum of single events. For example, student 1 in Group 1 was assessed 46 times and gathered information 503 times. Essentially, information-gathering events refer to pages containing texts or videos, and assessment events refer to pages where an evaluation of knowledge or skills is performed. Overall the student pilots in the first group had a ratio of about 9% of assessment for information gathering events, while the pilot students in the second group had a ratio of about 13%. The number of assessments includes repeated trials on assessment items. Given that the following sections focus on specific subsets of observations (ex. frequent sequences, or first attempt assessments only), Table 1 provides a high-level view and context for these learning events analysis.

Table 1. Distribution of assessment and information events for each student in the two groups.

Students	Assessment	Information	Total
Group 1			
1	46	503	549
2	45	497	542
3	51	514	565
4	42	495	537
5	52	477	529
6	49	512	561
7	47	547	594
8	57	478	535
Group 1 Total	389	4023	4412
Group 2			
a	42	305	347
b	55	323	378
c	37	259	296
d	34	280	314
e	41	311	352
f	37	284	321
Group 2 Total	246	1762	2008
Grand Total	635	5785	6420

4. SEQUENCE MINING

The objective of the application of sequence mining techniques to the learner dataset was to test the hypothesis that students who acted similarly in training would also perform similarly in the assessments. Results indicate that a significant relationship between students' behavioural patterns during training and performance on test problems exists.

For the analysis in this section, we utilized a data-driven approach to classify student activity and behaviour patterns in the web training courseware, with the purpose of identifying dependencies between the way students interact with the training material, and how the students perform on subsequent assessment-based tests and exercises. At a high level, the working hypothesis for this part of the study is thus that students who behave similarly (i.e. by exhibiting similar patterns of navigation activity when interacting with the courseware) will perform similarly in the assessments.

To test this hypothesis, we classified the students into two groups, using three different criteria: 1) those who scored above the median score on the assessments versus those who scored below the median, 2) those who scored above average on assessments versus those who scored below, and 3) classification according to response similarity. For this final classification scheme, we considered similarities in student success on a question-by-question basis. A distance function was introduced, with the distance between two students defined as the number of assessment questions for which one student gave the correct response and the other gave an incorrect response. K-means clustering was then used to divide the students into two groups in which in-class distances were minimized. Thus two students in the same class were likely to have scored the same (correct or incorrect) more often than two students in different classes. This particular analysis thus more closely strives to validate the working hypothesis that students who behave similarly will perform similarly in the assessments. So, rather than only judging similarity between two students only in terms of total score, we also took a view of how they scored in relation to each other in terms of the number of assessments in which both responded correctly or both responded incorrectly.

For each classification scheme above, the hypothesis is that students classified in the same group (i.e. those whose score similarly in assessments in terms of total score or response similarity) should have exhibited more similarities in how they interacted with the courseware during the learning phase. To test this, we utilized sequential pattern mining (using the SPAM [9] algorithm) to mine sequences of behaviour that were discriminative of each group (i.e. sequences of pages visited that were found to be highly frequent in one group and highly infrequent in the other), and then used leave-one-out cross-validation to test our ability to correctly classify each student based on the existence of these mined behavioural sequences.

Figure 2 shows the accuracy of our classifier for each classification scheme. For example, the leftmost bar indicates that we were able to correctly classify whether a student scored above or below the median score in 93% of the cases (as well as above/below average in 100% of cases and according to response similarity in 86% of cases), solely through analysis of behaviour patterns exhibited by the students when navigating through the courseware. The p-value for each statistic indicates the probability of achieving these results (or better) purely by chance. This indicates that a significant relationship exists between students' behavioural patterns during training and performance on test problems.

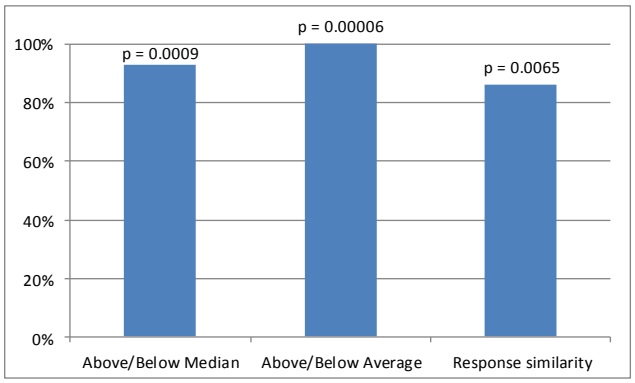


Figure 2. Results of sequence classification on students

To further examine the relationship between behaviour and results, we took a closer examination of the similarities between students when classified as either above or below average score, the scheme that was most successful in the test above. Here we generated the set of frequent behaviour patterns exhibited by each student, and then computed the Jaccard similarity of each pair by quantifying the degree of overlap in the set of frequent patterns for each student, where the Jaccard similarity of two sets A and B is equal to the size of the intersection of A and B, divided by the size of the union. Table 2 summarizes these results by showing, for each student, the average similarity to students who placed above and below the average. On average, students achieving a lower than average score had more similar behaviour to other students who achieved a lower than average score, and vice-versa. In fact, in all cases but one, each student behaved more similarly on average to students in its own group.

Table 2. Average similarity for each student to students with below/above average score

Below Average Students			Above Average Students		
Student	Similarity with below average students	Similarity with above average students	Student	Similarity with below average students	Similarity with above average students
1	0.125	0.080	3	0.059	0.071
2	0.078	0.068	4	0.100	0.075
5	0.047	0.033	a	0.051	0.068
6	0.070	0.061	b	0.063	0.112
7	0.032	0.026	c	0.024	0.042
8	0.127	0.075	d	0.063	0.133
			e	0.040	0.072
			f	0.059	0.142
Average	0.080	0.057		0.057	0.090

While there are wide-ranging behaviours that differentiate the two groups, Figures 3 and 4 point to two interesting behaviour patterns that were particularly prevalent in the initial dataset of 8 students. The first instance, in Figure 3, was highly frequent among the higher-achieving group, and quite infrequent among the lower-achieving group. This behaviour shows a lot of activity reviewing notes before completing a particular section and moving on. This could indicate that this note review had an impact on the success of the students. The second instance, in Figure 4, was highly frequent among the lower-achieving group, and quite infrequent among the higher-achieving group. This behaviour shows a lot of activity around calculations regarding take-off. This could provide

a clue into where the less successful students are going wrong, and thus where improvements to the courseware may be made.

```
1.  Review_Introduction_1, Review_Introduction_2,
2.  Full_Review_Notes_Mission_Planning_1,
3.  Full_Review_Notes_Landing_Limits_and_Procedures_2,
4.  Full_Review_Notes_Landing_Crosswinds_3,
5.  Full_Review_Notes_Takeoff_Procedure_4,
6.  Full_Review_Notes_Takeoff_Conditions_5,
7.  Full_Review_Notes_Takeoff_Crosswinds_6,
8.  Full_Review_Notes_Landing_Calculations_7,
9.  Full_Review_Notes_Takeoff_Calculations_8,
10. Full_Review_Notes_ControlUnit_Invalid_9,
11. Full_Review_Notes_ControlUnit_Calculations_10,
12. Transition_To_Test-GUI_MAP,
13. Lesson_Conclusion_Pass
```

Figure 3. Example behaviour of the higher-performing group

```
1.  Select_Calculation-Takeoff_Crosswinds_1-
2.  Select_Calculation-Takeoff_Pitch_1-Takeoff_Pitch_2,
3.  GUI_MAP-Calculations_Introduction_1-
    Calculations_Introduction_2-
    Calculations_Introduction_3- Invalid_11-Invalid_12,
4.  Invalid_14-How_To_Use_Introduction_1-
    How_To_Use_Introduction_2,
```

Figure 4. Example behaviour of the lower-performing group

This result has a number of implications. First, it demonstrates a tangible correlation between how students choose to navigate the courseware and how well they perform on assessments. Second, it establishes clear evidence that opportunities exist to predict student achievement during the learning phase, when remedial action can be taken to improve comprehension. Finally, the ability to identify the key behaviours that have the highest impact on how a student will perform can facilitate strategic managerial decision making on how to direct the flow of student activity through the courseware.

5. ADDITIVE FACTOR MODELS

The Additive Factor Model (AFM) was chosen because it represents a common technique in educational data mining [12]. By using this data analysis technique, we were seeking estimations for parameters for student proficiencies, as well as items difficulty, and competencies easiness. AFM is a model for assessing the quality of an items-to-skills mapping, based on its ability to predict empirical observations of student results [10]. It may be seen as a generalization of Item Response Theory [11], where the response depends not only on item difficulty and student proficiency, but also on underlying knowledge components (KC) and the sequence in which they are met. In AFM, these knowledge components can be associated with competencies, skills, or declarative knowledge that are responsible for a student's performance. The mapping between an item (question, task, problem) and knowledge components is provided in the form of a binary Q-matrix $\mathbf{Q}=[q_{ik}]$, where $q_{ik}=1$ indicates that item i is associated to knowledge component k [13]. The probability that a student j will correctly answer an item i is modelled using a mixed-effect logistic regression

$$P(Y_{ij}=1|\alpha,\beta,\gamma)=\frac{1}{1+\exp\left(-(\alpha_j+\sum_k\beta_k q_{ik}+\sum_k\gamma_k q_{ik}t_{jk})\right)} \qquad (1)$$

where α_j is the proficiency of student j (higher proficiency yields higher success rate), β_k is the easiness and γ_k the learning rate for knowledge component k (higher easiness yields higher success, higher learning rate means increased success on subsequent trials)[2]. The observed student sequence is summarized in the opportunity t_{jk}, i.e. the number of times student j has met knowledge component k. As learning progresses, increasing opportunity translates into higher probability of success in items associated with that KC.

Our learner dataset contains 38 items, taken by 14 students (in two sessions of eight and six) between zero and four times each, resulting in 533 transactions.[3] The course designers provided the Q-matrix mapping the 38 items to 14 knowledge components (Figure 5, where the items are the specific questions or problems that the students had to answer or solve, while the knowledge components are the underlying knowledge and skills accounting for the learner's performance on those questions or problems.

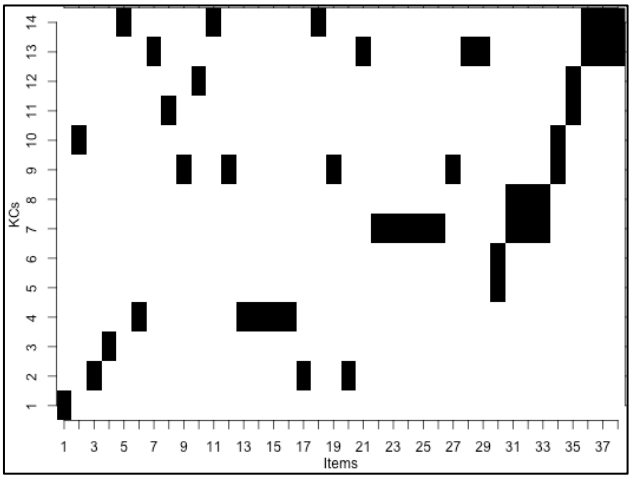


Figure 5: Q-matrix from courseware designer: 38 items x14 KCs.

Estimation of the AFM model parameters is done by maximizing the likelihood[4] on the transactions, with the constraint that learning rates are kept positive, and a slight regularization on the alpha parameters in order to keep them within the [-3; 3] range.

5.1 Student Proficiency

We analyse the proficiency of the two groups of students using the estimated alpha parameters. Figure 6 shows that the first group of students (1-8) has overall a lower proficiency than the second group (a-f). The two students with lower proficiency in the second group (b and c) have estimated proficiencies on par with the best two students from the first group (3 and 4). Student 5 clearly displays the lowest proficiency by far.

This is partly reflected in the observed success rates, which range from 58.5% for student 5, to 100% for student d. We learned *post analysis* that the second group had received an improved set of instructions. Although there was no difference between the first and second groups in expectations, motivation or engagement with the training material, the improved instructions have a clear

[2] Proficiency and easiness values are relative to the other values in the set, and should not be interpreted as actual success rates.

[3] Each transaction records one student's result on one item.

[4] We use a conjugate gradient algorithm. Any optimization method would work similarly as the log-likelihood is convex.

impact on the estimated proficiency for the second group. This validates the effectiveness of the change.

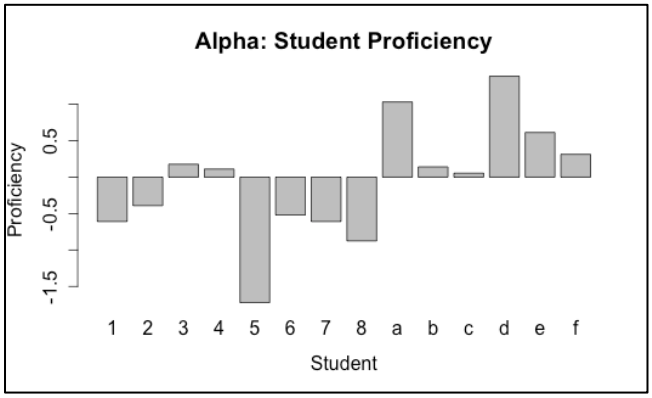


Figure 6: Student proficiency, estimated by AFM.

5.2 Competency Analysis

We analyse the competencies through the estimated beta and gamma parameters. Note that the actual parameter values are difficult to interpret separately, as various combinations of beta, gamma and opportunity may yield similar probabilities (Eq. 1). They do make sense in combination of the base "easiness" beta and learning rate gamma, to explain how the probability of success changes as the number of opportunity increases. As a consequence, rather than looking at actual parameter values, we relate them to the corresponding prediction ability. We analyse competencies by looking at the probability to fail on items associated by each knowledge component on the first three opportunities, for a hypothetical student with a proficiency parameter of zero. Figure 7 shows this for 11 knowledge components (The easiest KCs, 1, 4 and 11, get 0% for both predicted and observed error from the first attempts).

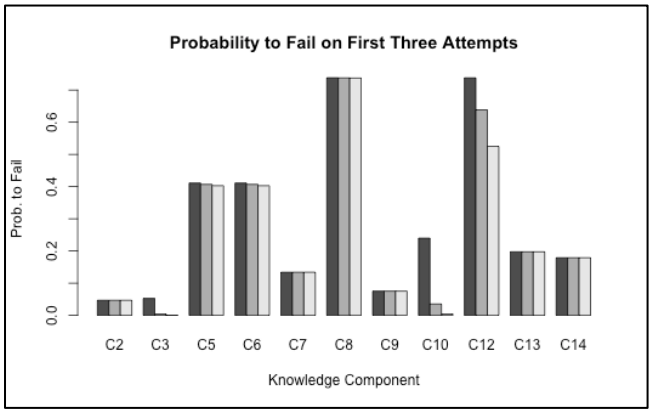


Figure 7: Probability of error for several knowledge components.

Note that due to the constraint that the learning rate is positive the probability to fail is always decreasing (Eq. 1). Learning is clearly apparent for several competencies (C3, C10 and C12), as shown by the clear drop in probability to fail as the KC is addressed. For C5 and C6, learning is much slower, and the error rate stays around 41%. However, this observation should be mitigated by the fact that these knowledge components are only associated with one item and always together (Figure 5). There is therefore very little data to estimate learning on these competencies, as most students took that item only once. When considered in combination in item #30, KCs C5 and C6 yield a predicted error on this item of 36%. In addition, this points to a possible refinement of the Q-matrix: these two knowledge components could be merged with no loss of modelling capacity.

Probability of failure seems consistently high for C8. However, Figure 5 shows that this knowledge component always appear together with C7 (which also appears alone). Due to the additive nature of the AFM model, the actual probability of success for items featuring C8 actually combine the easiness and learning rates for both C7 and C8, resulting in a probability of failure of 30.3%. Items involving both C7 and C8 are significantly harder than items involving C7 alone, and the AFM model adjusts for this fact by estimating a low easiness (high difficulty) for knowledge component C8.

The analysis of the AFM results therefore provides us with non-trivial insight into 1) the proficiency of the students taking the course, and 2) the difficulty and learning rates of the various competencies addressed in the course. It also suggests possible refinements of the competency framework produced by the course designer. Finally, despite the clear difference between the two groups of students, we have also observed that the estimates for the parameters related to competencies (β_k and γ_k) are consistent across the two groups.

6. CONCLUSION

To address the challenges of pilot training in the early 2000s, civil aviation stakeholders like CASA, ICAO, and concurrently the USAF have been promoting competency-based training as a training model. In addition to focusing on competencies rather than hours, the industry has also brought to bear actual flight monitoring data as a source to determine learning objectives. The essential element evidence-based training introduces to competency based-training is the reference to operational data as a means to identify key competencies, in addition to the analysis performed by subject matter experts. A literature review also revealed that a combination of competency, evidence, and scenario-based training approaches can form the basis for the next generation of pilot training system. The latter approach being consistent with the principles of situated learning theory, which argues that learning best takes place in the context in which it is going to be used. The paper focused essentially on the assessment of knowledge components using sequence mining and logistic regression for the purpose of understanding learning processes and improving learning scenarios. The data used for these analyses was collected in the context of pilot training using a scenario-based approach for reviewing basic landing and taking off flight operations.

The objective of the application of sequence mining techniques to the learner dataset was to test the hypothesis that students who acted similarly in training would also perform similarly in the assessments. Results indicate that a significant relationship between students' behavioural patterns during training and performance on test problems exists.

The Additive Factor Model, a model for assessing the quality of an items-to-skills mapping based on empirical observations of student results, was used to estimate student proficiency and knowledge components difficulty. Our analysis indicated a clear difference between students from two groups in the data. It also helped us identify competencies that are inherently easy, as well as hard competencies for which learning allows the probability of failure to quickly drop over subsequent attempts. It also suggests changes in the competency framework in which knowledge components could be merged with no loss of modelling capacity.

Together, the application of the descriptive method of sequence mining, and the predictive technique of additive factor models, provide results that may be used to evaluate and improve instructional design.

Some potential future directions for the project include: a) collecting more data, using the same approach for additional data sets, and comparing the result; b) developing alternative methods, and using the methods on same data sets to test and compare results; and c) conducting validation with instructional design experts in the relevant domain.

7. ACKNOWLEDGMENTS

The NRC project team would like to thank Dr. Susanne Lajoie (McGill University), who helped the NRC team to obtain its ethics certificate by providing the relevant documentation supporting McGill's ethics request to process CAE pilot learning data. The authors would also like to thank the following reviewers from CAE: Paula Mazzaferro, David Graham, and Graham Estey.

8. REFERENCES

[1] International Civil Aviation Organization, *Manual of Evidence-based Training*, First edit. Montreal, Canada: International Civil Aviation Organization, 2013.

[2] Civil Aviation Safety Authority, "Competency Based Training and Assessment in the Aviation Environment," 2009.

[3] C. M. Colegrove and G. M. Alliger, "Mission Essential Competencies : Defining Combat Mission Readiness in a Novel Way," in *RTO SAS Symposium on "Air Mission Training Through Distributed Simulation (MTDS) Achieving and Maintaining Readiness,"* 2002, vol. 323, p. 22.

[4] International Civil Aviation Organization, *Quality Assurance Manual for Flight Procedure Design. Flight Validation Pilot Training and Evaluation (Development of a Flight Validation Pilot Training Programme)*, First edit., vol. 6. Montreal, Canada: International Civil Aviation Organization, 2012.

[5] R. Chapman and C. Colegrove, "Transforming operational training in the Combat Air Forces.," *Mil. Psychol.*, vol. 25, no. 3, pp. 177–190, 2013.

[6] Air and Space Academy, "Dealing with Unforeseen Situations in Flight," Bruguières, France, 2013.

[7] M. Thurber, "The future of pilot training," *Aviation International News Online*, 2014. [Online]. Available: http://www.flightresearch.com/pdfs/AIN_Pg_20-28.pdf. [Accessed: 01-Feb-2015].

[8] J. MacMillan, E. B. Entin, R. Morley, and W. Bennett, "Measuring team performance in complex and dynamic military environments: The SPOTLITE method.," *Mil. Psychol.*, vol. 25, no. 3, pp. 266–279, 2013.

[9] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 429–435.

[10] H. Cen, "Generalized Learning Factors Analysis: Improving cognitive Models with Machine Learning," Carnegie Mellon University, 2009.

[11] R. D. Bock, "A Brief History of Item Theory Response," *Educ. Meas. Issues Pract.*, vol. 16, no. 4, pp. 21–33, 1997.

[12] A. Pena-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

[13] K. K. Tatsuoka, "Rule space: an approach for dealing with misconceptions based on item response theory," *J. Educ. Meas.*, vol. 20, no. 4, pp. 345–354, 1983.

A Scalable Learning Analytics Platform for Automated Writing Feedback

Jacqueline Feild
McGraw-Hill Education
Boston, MA
jacqueline.feild
@mheducation.com

Nicholas Lewkow
McGraw-Hill Education
Boston, MA
nicholas.lewkow
@mheducation.com

Neil Zimmerman
McGraw-Hill Education
Boston, MA
neil.zimmerman
@mheducation.com

David Boulanger
Athabasca University
Edmonton, CA
david.boulanger
@dbu.onmicrosoft.com

Jeremie Seanosky
Athabasca University
Edmonton, CA
jeremie
@rsdv.ca

ABSTRACT

In this paper, we describe a scalable learning analytics platform which runs generalized analytics models on educational data in parallel. As a proof of concept, we use this platform as a base for an end-to-end automated writing feedback system. The system allows students to view feedback on their writing in near real-time, edit their writing based on the feedback provided, and observe the progression of their performance over time. Providing students with detailed feedback is an important part of improving writing skills and an essential component towards solving Bloom's "two sigma" problem in education.

We evaluate our feedback system in two ways. First, we evaluate the effectiveness of the feedback for students with an ongoing pilot study with eight hundred students who are using the learning analytics platform in a college English course. In addition, we process an existing set of graded student essays and analyze the performance feedback. Results show a correlation between feedback values and human graded scores.

Keywords

Analytic Tools for Learners; Automated Essay Feedback; Scalable Analytics; Performance Feedback; Natural Language Processing

1. INTRODUCTION

Performance feedback is essential for self-regulated learning, which is an attribute of highly effective learners [3, 18]. Bloom has shown that providing formative feedback to students increases performance, compared to only providing final feedback [1]. This allows students to develop and implement actionable strategies for improving performance as they progress. Formative feedback is even more effective if it can be given in near real-time [7, 13].

In this paper we describe a scalable platform for learning analytics called OpenACRE (Analytics Collaborative Research Environment) which is currently in development to be released as open source. OpenACRE allows for ingestion of heterogeneous educational data from multiple source systems, long-term storage of raw data, running arbitrary models on the raw data using a parallel analytics engine, and short-term storage of resulting analytics for use by students, teachers, and researchers. As a proof of concept, we implement an end-to-end writing feedback system utilizing OpenACRE. Writing feedback is especially hard to provide in real-time and at scale as it is computationally expensive, making it well suited for the capabilities provided by OpenACRE.

There are several other existing writing feedback systems which provide various feedback to students, for example Revision Assistant, WriteToLearn, and Writing Pal [17, 14, 12]. While these systems provide useful information, they are either commercial black boxes which do not allow for modification, or are intelligent tutoring systems which provide writing instruction through customized modules. OpenACRE stands apart by providing the ability to develop and deploy new analytical models at scale, making it useful for researchers to test new feedback algorithms, predictive models, or reporting dashboards on a large number of students.

To evaluate our proof of concept system in a classroom setting, an efficacy study is currently underway to investigate the usefulness of the feedback to improve student performance. The study consists of eight hundred college students who are learning English at VNR VJIET in India. Additionally, we evaluate the feedback from 13,000 existing student essays and compare it to the human graded scores.

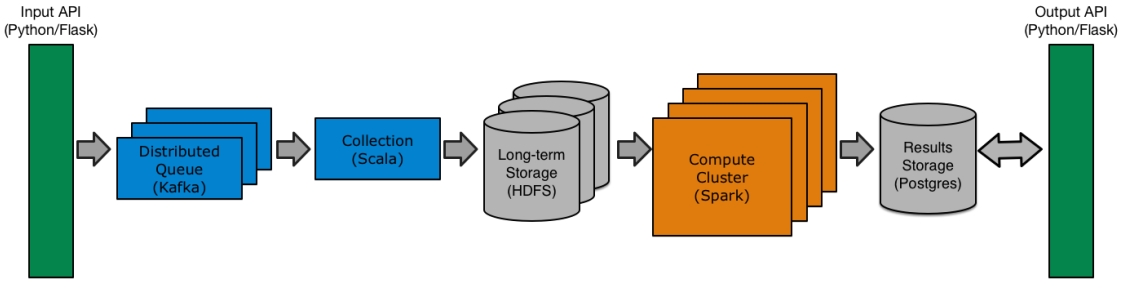


Figure 1: Architecture diagram of the learning analytics platform, corresponding to the middle box in Figure 2. Data is ingested by the input API and placed into a distributed queueing system which is implemented using Kafka. A collection service, implemented in Scala, pulls data from the queue and stores it in long-term storage, which is implemented using Hadoop Distributed File System (HDFS). The compute cluster runs models in parallel on the data in long-term storage and persists output views to the results store, implemented in PostgreSQL. Output views can then be accessed through the output API. Both the input and output APIs are RESTful and implemented in Python using Flask.

2. OPENACRE

The OpenACRE platform consists of an input and output API, long- and short-term databases, and a parallel computation cluster. A low-level diagram is shown in Figure 1. This platform is designed to handle the challenges of scalability, resiliency to data loss, and fault tolerance. Additionally, OpenACRE is built to be extensible for future models, without the need for drastic modification to the system as a whole. For example, models which perform machine learning algorithms, complex aggregations, and graph analysis could all be implemented to run on OpenACRE. These models could include traditional classroom statistics, score predictions, or personalized learning recommendations.

Learning event data is ingested into OpenACRE through the input API and persisted to the long-term data store. The input API for OpenACRE is implemented in a RESTful fashion using Python with the Flask package. RESTful APIs are used because they are stateless, easily extended for future functionality, and agnostic to programming language. The input API accepts event data from external sources and temporarily stores the events in a queueing system. We utilized open source Apache Kafka for our queueing system as it is distributed, durable, and supports APIs in several commonly used languages. Next, a collection service takes events from the queue and stores them in a long term data store. Here we use the open source Hadoop Distributed File System (HDFS) since it is distributed and fault tolerant. The collection service in OpenACRE is written in Scala, but any language supported by the Kafka and Hadoop APIs could also be used. The event data stored in HDFS is kept in its original "raw" form and is never altered. Storing unaltered event data allows for arbitrary computation and the implementation of future models without knowledge of those models beforehand.

Next, the computation engine runs analytical models by taking data from the long term store and performing transformations/aggregations to create new output views. These output views can be accessed by users through the output API. Open source Apache Spark was used for our computation engine as it allows for user-friendly parallel computation, horizontal scalability on commodity hardware, and contains a rich set of APIs ranging from simple map-reduce to machine learning algorithms. Additionally, Apache Spark currently implements APIs in Java, Python, and Scala.

Output views from a given model are written to the results store database which is implemented using PostgreSQL in OpenACRE. PostgreSQL was used as it is open source, has APIs in several languages, and provides a familiar SQL interface for queries. From the results store, output views are provided to external users through the output API. Similar to the input API, this is implemented as a RESTful API so it is stateless and can be easily accessed from the majority of modern languages. The output API can then accessed by other backend systems or user facing systems, such as dashboards. The combination of all the OpenACRE components listed above results in a learning analytics platform which can ingest arbitrary learning event data, apply parallel analytic models to the data, and provide the results of the analytics to external systems and dashboards in a generic fashion.

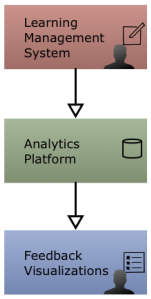


Figure 2: High-level diagram of the end-to-end writing feedback system. The learning management system and feedback visualizations are student-facing while the learning analytics platform stores writing data and computes feedback.

While any type of data format could be ingested into OpenACRE, we chose the standardized learning event format called Caliper, supported by IMS Global [4]. Caliper defines a set of standard learning events composed of actor-action-object triples. An example event is 'student-submits-quiz1'. While actor-action-object triples are also used in other standardized learning event formats like TinCan [16], Caliper has a significant benefit in that is uses JSON-LD, which is a schema-based JSON format. In addition to being schema-based, JSON-LD allows for easy mappings from JSON to domain-specific ontologies.

3. END-TO-END WRITING FEEDBACK SYSTEM

As a proof of concept, we built an end-to-end writing feedback system with OpenACRE at the core. Writing feedback is an excellent use case for OpenACRE as it is very computationally expensive, requiring approximately 12 seconds per essay for our feedback model. This large processing time results in almost 7 days of computation for a single assignment in a large MOOC of 50,000 students. Implementing our writing feedback model on OpenACRE allows that computation time to be cut to hours or minutes, depending on the size of the computation cluster.

The end-to-end proof of concept system includes the student facing system, which collects student writing data from their learning management system (LMS) and displays the automated feedback visualizations, and the backend system built on OpenACRE, which stores and analyzes the student data. Figure 2 shows a high-level view of this system, including both the student facing and backend systems.

The typical workflow for a student using this system includes:

1. Log in to writing course using an LMS
2. Start a writing assignment
3. Save the writing assignment
4. View visualizations of writing feedback
5. Edit writing assignment based on provided feedback
6. Save the writing assignment
7. Repeat steps 4-6 as needed
8. Submit assignment

This workflow provides feedback to students at regular intervals and gives students the opportunity to improve their writing before submitting their assignment. The ongoing pilot provides feedback in 24 hour increments due to cost constraints on the size of the computation cluster. Since the LMS which students are using is instrumented to directly collect writing data, there is no need to use an additional feedback system. This allows for an intuitive interaction between the student and their LMS, while collecting data for feedback at the same time. In our implementation, we utilized Moodle for our LMS as it is open source, familiar to both students and educators, and was easily instrumented to collect writing data as Caliper events and send those events to OpenACRE.

We designed a custom dashboard to display feedback visualizations to students and instructors. These include both a snapshot of overall feedback and the progression of feedback over time.

3.1 Feedback Competences

The feedback provided by our system is composed of seventeen writing competences which have been developed over the last several years [9]. These include traditional writing metrics such as spelling and grammatical accuracy as well as more advanced metrics that capture sentiment and writing flow. In the following sections, we describe several groups of writing metrics and define the competences we implement within them.

3.1.1 Traditional Metrics

Traditional writing metrics include competences that are often used by teachers to evaluate student writing. The competences implemented in our system from this category include vocabulary, spelling, grammatical accuracy, and lexical diversity. The vocabulary competence represents the amount of unique words in the student's text. As the student uses more unique words in their writing, the vocabulary competence increases. The spelling competence measures the percentage of incorrectly spelled words used. This competence increases as the percentage of misspelled words in the text decreases. Similar to spelling, the grammatical accuracy competence measures the percentage of grammatical errors in the text. This competence value increases as the percentage of grammatical errors decreases. Finally, the lexical diversity competence measures the percentage of unique words in the text. The value increases as students use more unique words relative to the size of the text.

3.1.2 Advanced Metrics

Advanced writing metrics highlight more subtle and complex characteristics of English writing. While not always explicitly listed in a writing rubric, these metrics are important for proficient English writing. The competences implemented in our system from this category include modifier complexity, noun phrase complexity, and tense agreement. The modifier complexity competence represents the amount of noun or verb modifiers which are used in the student's text. A high number of noun or verb modifiers indicates that the writing is more complex and expressive. The noun phrase complexity competence analyzes the number of noun phrases in the student's text. This metric attempts to measure the linguistic complexity for a piece of writing, as more noun phrases typically indicates richer sentences. Finally, the tense agreement competence measures the consistency of verb conjugations in the text. This competence value increases when verbs are conjugated consistently throughout a piece of writing.

3.1.3 Flow Metrics

Writing flow metrics measure how ideas are connected both within adjacent sentences and throughout entire pieces of text. The competences we implement in this category are

local cohesion, global cohesion, and connectivity. Local cohesion tracks the flow of ideas from sentence to sentence. Writing that contains adjacent sentences with similar nouns and verbs receives a higher local cohesion score. Similarly, global cohesion tracks the flow of ideas throughout an entire piece of writing, which is also measured by the similarity of nouns and verbs throughout the text. Connectivity measures the use of phrases that connect ideas to one another. Text with more coordinating conjunctions receives a higher connectivity score.

3.1.4 Descriptive Metrics

These writing metrics measure how descriptive a piece of writing is in several different ways. The competences we implement in this category are concreteness, imagery, familiarity, and conciseness. Concreteness measures the degree to which the text refers to tangible objects. Higher concreteness scores are obtained by using more words that refer to tangible objects. Imagery gives a measure of the amount of words within the text which evoke a mental image. Similarly, familiarity measures the amount of words in a text that are commonly used. The calculation of concreteness, imagery and familiarity are based on pre-defined scores in each category for commonly used words. These pre-defined scores were determined experimentally by asking human subjects to rate words in these three categories [5]. Conciseness measures the ratio of content words in the text. Writing that includes more nouns, verbs, adverbs and adjectives receives a higher conciseness score.

3.1.5 Sentiment Metrics

Sentiment metrics reflect the tone or feel of a piece of writing. These are computed using state-of-the-art techniques with the Stanford CoreNLP library [10, 15]. The required sentiment may vary based on the type of writing or subject matter. The competences we implement in this category include negative tone, neutral tone, and positive tone. The negative tone competence describes the degree to which the writing exhibits negative sentiment. Similarly, the neutral and positive tone competences describe the degree to which the writing exhibits neutral or positive sentiment. All three of these competences measure the amount of negative, neutral, or positive words in the writing.

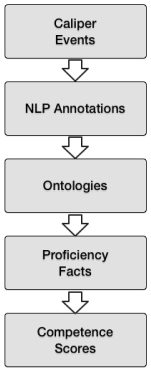


Figure 3: High-level diagram showing the flow of the openSCALE algorithm from caliper events to competence scores.

3.2 OpenSCALE

The analytics model implemented in this automatic writing feedback system is called OpenSCALE [2]. This model parses text with Stanford CoreNLP library [10], creates ontologies and facts from the annotated text, and aggregates the facts into competence scores for students.

A high-level view of the transformations which go from text to competence scores is described in Figure 3. First, the text is annotated using the Stanford CoreNLP library [10]. The annotations include tokenization of the text into words and sentences, part of speech tagging, syntactic parsing and sentiment analysis. These annotations are used to create an ontology of the relationships between words, sentences and paragraphs in the text, including both their structure and semantic meaning. For each piece of text, openSCALE creates one ontology using the open source Apache Jena library.

Next, each ontology is put through an inferencing layer, which looks for patterns in the ontology that show evidence of students having a particular skill/competence and creates proficiency facts. Each fact includes information about the degree of competence (weight) for a unique student-assignment attempt-time. Many facts are generated from a single ontology going through the inferencing layer. The inferencing layer in openSCALE is implemented using VIStology's BaseVISor framework [11]. BaseVISor works by passing a set of rules dictating how facts are generated for a given ontology. The ability for users to specify specific rules allows for great flexibility as different instructors could potentially dictate what is seen as evidence of different skills/competences. The current implementation of openSCALE uses a default set of rules which are used by BaseVISor.

Finally, the proficiency facts are aggregated to generate final scores for each competence. The main flow of the fact aggregations for student, competence, assignment attempt, and time is:

1. Sum the weights for all facts with the same student-competence-assignment attempt-time
2. For each fact F (student S - assignment attempt A - competence C - time T):
   (a) Find all facts at or before time T with the student S - competence C
   (b) Keep the facts of the newest attempt for each assignment
   (c) Sum the competence weights and update F

The final, aggregated facts are used to generate the competence progression view in the results store. The view displaying a snapshot of overall feedback is created by taking the latest aggregated facts for each competence.

4. PILOT RESEARCH STUDY

We are currently running a pilot research study to test the usefulness of the feedback system for increasing student writing performance. Eight hundred first year engineering students at VNR VJIET in India are using our system to complete up to twenty writing assignments.

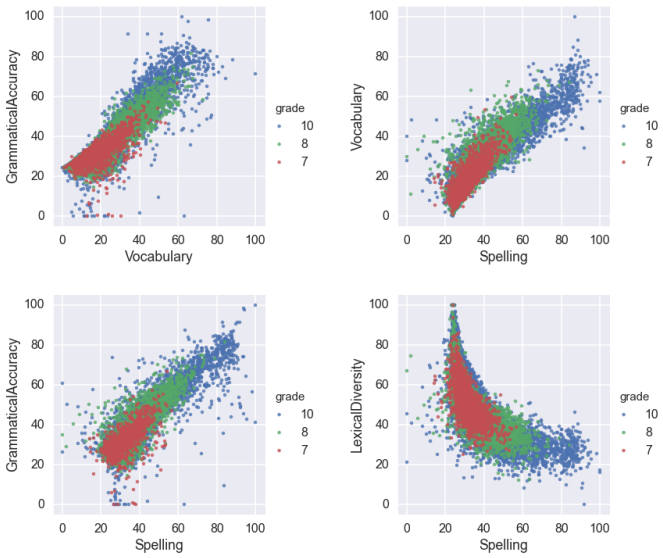


Figure 4: Scatter plots showing two competence scores plotted against each other all essays in the dataset. Data points are colored to distinguish essays from 7th, 8th, and 10th grades.

The current pilot study is an observational study and will use the method of propensity score analysis to determine the effectiveness of the feedback visualizations [6]. Students will also fill out surveys about the feedback they received and its usefulness.

5. ANALYSIS WITH EXAMPLE STUDENT ESSAYS

While the pilot is in progress, to additionally evaluate the usefulness of the writing feedback system, feedback was generated from a dataset containing about 13,000 anonymized student essays which have been graded by humans. The dataset was obtained from the Kaggle competition for automated essay scoring [8] and includes essays for students in 7th, 8th, and 10th grade. A total of eight different groups of essays are contained within the dataset, each with a different writing prompt and grading rubric. For our experiments, all essays were mixed together, grouped only by grade of the student, and all human grades have been normalized to range between 0-100.

First, we investigated correlations between competence types. Figure 4 shows competence vs competence scatter plots for grammatical accuracy, vocabulary, spelling, and lexical diversity. Data points are colored to distinguish between 7th, 8th, and 10th grade essays. Strong linear relationships can be seen for both plots containing grammatical accuracy in addition to vocabulary vs spelling. Additionally, an interesting relationship between lexical diversity and spelling can be seen in Figure 4. This plot shows that no students have high values in both lexical diversity and spelling simultaneously. To achieve high scores in the spelling competence, a longer essay is required with the majority of the words spelled correctly. In contrast, long essays tend to have lower lexical diversity competence values as more words are repeated in longer writings. The resulting balance of these two competences can be clearly seen in Figure 4.

Next, we plotted competence values against human graded scores. Figure 5 shows competence values for connectivity, grammatical accuracy, modifier complexity, and noun phrase complexity plotted against the graded score. Connectivity, grammatical accuracy, and noun phrase complexity all show the trend that increased competence values correlate to higher graded scores. The plot displaying modifier complexity shows the graded score initially increasing with competence value. There is a point which this trend stops and the average score stays constant, or even decreases, as the competence value increases. This data suggests that essays with a lot of complex modifier usage score the same or even lower than corresponding essays with moderate modifier usage. The above analysis gives us confidence in the usefulness of the competence feedback for improving performance.

6. CONCLUSIONS

Providing real-time feedback to students is an important component to solving Bloom's two sigma problem. In this paper we described a scalable learning analytics platform (OpenACRE) which is able to ingest educational data from multiple external systems and provide analytics on that data in near real-time. We demonstrated the usefulness of this platform with the implementation of a writing feedback system and are currently running a pilot research study to evaluate its effectiveness with eight hundred first-year engineering students at a university in India. We also showed that competence values correlated with human graded scores on a set of existing student essays. Development is currently underway to release OpenACRE as an open source project for other educational researchers.

7. ACKNOWLEDGMENTS

This paper is based on work supported by the McGraw-Hill Education Digital Platform Group (MHE DPG). Despite provided support, any opinions, findings, conclusions

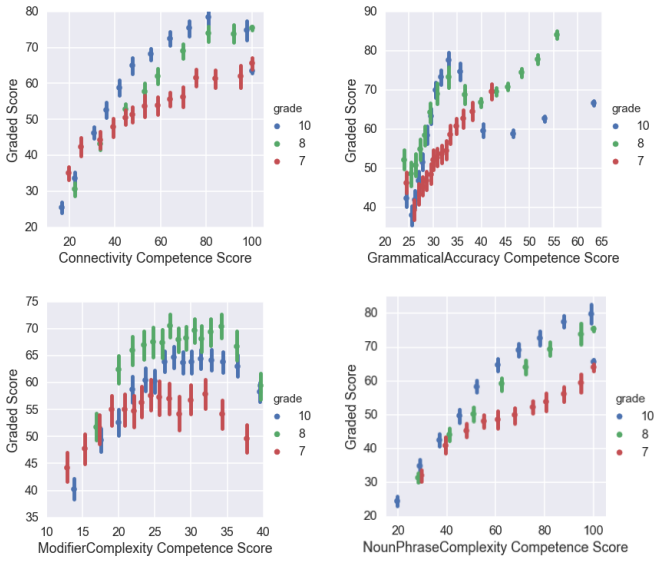


Figure 5: Average essay score as a function of several competence values. Error bars display standard deviation from the mean score. Data points are colored to distinguish essays from 7th, 8th, and 10th grades.

or recommendations expressed in this paper are those of the authors and do not necessarily reflect positions or policies of the company.

8. ADDITIONAL AUTHORS

Additional authors: Mark Riedesel (McGraw-Hill Education, Boston MA, email: `mark.riedesel@mheducation.com`), Alfred Essa (McGraw-Hill Education, Boston MA, email: `alfred.essa@mheducation.com`), Vive Kumar (Athabasca University, Edmonton CA, email: `vive@athabascau.ca`), Kinshuk (Athabasca University, Edmonton CA, email: `kinshuk@athabascau.ca`) and Sandhya Kode (IIIT Hyderabad, Hyderabad, India, email: `sandhya.kode@gmail.com`).

9. REFERENCES

[1] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, pages 4–16, 1984.

[2] D. Boulanger et al. Scale: A competence analytics framework. In *State-of-the-Art and Future Directions of Smart Learning*, pages 19–30. Springer, 2016.

[3] D. L. Butler and P. H. Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, 65(3):245–281, 1995.

[4] I. G. L. Consortium et al. Learning measurement for analytics whitepaper, 2013.

[5] K. J. Gilhooly and R. H. Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427, 1980.

[6] S. Guo and M. W. Fraser. Propensity score analysis. *Statistical methods and applications*, 12, 2015.

[7] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[8] Kaggle. The hewlett foundation: Automated essay scoring. https://www.kaggle.com/c/asap-aes, 2012. Accessed: 2015-10-01.

[9] V. Kumar et al. Mobile computing and mixed-initiative support for writing competence. *Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers: Technology Enhanced Support for Learners and Teachers*, page 327, 2011.

[10] C. D. Manning et al. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[11] C. J. Matheus et al. Basevisor: A triples-based inference engine outfitted to process ruleml and r-entailment rules. In *Rules and Rule Markup Languages for the Semantic Web, Second International Conference on*, pages 67–74. IEEE, 2006.

[12] D. S. McNamara et al. The writing-pal: Natural language algorithms to support intelligent tutoring on writing strategies. *Applied natural language processing and content analysis: Identification, investigation, and resolution*, pages 298–311, 2012.

[13] D. J. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.

[14] Pearson. The research behind writetolearn, 2007.

[15] R. Socher et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[16] R. Software. Tin can api. https://tincanapi.com, 2015.

[17] turnitin. Turnitin revision assistant, 2015.

[18] B. J. Zimmerman. Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1):3–17, 1990.

An Automated Test of Motor Skills for Job Selection and Feedback

Bhanu Pratap Singh
Aspiring Minds
bhanu.pratap@aspiringminds.com

Varun Aggarwal
Aspiring Minds
varun@aspiringminds.com

ABSTRACT

Motor skills are required in a large number of blue collar jobs today. However, no automated means exist to test and provide feedback on these skills. In this paper, we explore the use of touch-screen surfaces and tablet-apps to measure these skills. We design novel app-based gamified-tests to measure one's motor skills. We show this information to strongly predict the job performance of skilled workers in three different occupational roles. The results presented in this work make a strong case for using such automated, touch-screen based tests in job selection and to provide automatic feedback. To the best of the authors' knowledge, this is the first attempt at using touch-screen devices to scalably and reliably measure motor skills.

Keywords

Motor skills; Touch-screen devices; Tablets; Assessments; Blue collar jobs.

1. INTRODUCTION

There are many standardized automated tests of language, knowledge, cognitive skills and personality [8, 1, 2]. These tests, often taken on a computer, are good predictors of academic achievement and job performance in the knowledge economy. They have also enabled automated feedback and credentials for learners.

We are interested in automating assessments of motor skills required for vocational jobs such as tailoring, plumbing and carpentry. In the Occupational Information Network (O*NET) database of job descriptions [11], 350 out of 1,065 jobs need moderate to high motor skills. There has been tremendous interest worldwide among employers and professional organizations in training and efficiently identifying people that possess the skills for such hands-on occupations [3, 9]. There have been several validated, non-automated tests like the Purdue Pegboard test [13] and the O'Connor Tweezer Dexterity test [12]. However, no serious attempt has been made to develop and validate automated tests for this purpose. Automated assessments so far have exploited the power of PCs and laptops. We wish to make use of a touch interface, in the form of tablet devices, to test motor skills.

The ability to test motor skills automatically using touch interfaces would allow it to scale extremely well, given the high market penetration of inexpensive tablet devices in the last five years. This would enable people to measure their motor skills right from their homes and receive feedback toward self-improvement. There is substantial evidence that motor skills among adults can be improved [14] and that explicit motor skills feedback and instructions help do so [7, 5, 10]. Also, test takers can learn how suitable they are for a given job, get credentials for the skills they have acquired and apply for jobs that are the best match for their particular skill sets. Companies, for their part, can remotely administer these tests and can use the scores registered and the certificates offered to find a quality workforce, making the identification of suitable candidates easy, cheap, and scalable. This has the potential to make the blue-collar labor market considerably more efficient, similar to the effect automated testing has had on the white-collar labor market.

We apply the classical procedure used in developing skill assessments to develop tests which measure motor skills. We first identify the skills that are most useful to test. We then develop app-based tests that run on tablets and have the potential to measure these skills.[1] We use capacitive touch interfaces in this work, which are very popular these days. The app-based tests are designed in such a way that they *exercise* the motor skills of a person and are of varying degrees of difficulty. Candidates undergo testing through various movements of their fingers, hands and arms. We develop scores for each app based on the test taker's interaction with it. We then test whether these scores are predictive of job/task performance in three occupational roles: tailors, machinists/grinders and machine operators. If our test scores can indeed predict performance in job roles, they could be useful both to provide corporations with a way to filter/evaluate candidates for such jobs and to give feedback to job seekers and those interested in training for such specific fields.

We found that the app-based test scores can predict job performance across multiple parameters that are considered in

[1]We consider tablets instead of smartphones to assess wider movements of arms and shoulders.

evaluating the three job roles enumerated above. The correlation values range from 0.19 to 0.38. These are comparable to, and in cases outperform, those reported historically for manual motor skill tests in predicting job performance ($0.06 - 0.30$, Table 1). This provides strong support for the use of automated touch-screen tests for measuring motor skills for job selection and recruitment. The paper makes the following contributions:

- It is the first attempt to design a touch-screen based test of motor skills. We design a number of novel apps for this purpose.
- We show that there is firm supporting evidence for using app-based scores in the job selection/recruitment process for multiple jobs. This can yield tremendous scalability in the process of hiring blue-collar workers and providing them feedback.

This paper is organized as follows: §2 discusses the motor skills we measure; §3 discusses the design of our apps; §4 lays out the experiment objective and analyzes our results and finally, §5 concludes the paper.

2. MOTOR SKILLS TO MEASURE

We wished to identify motor skills that predict job performance for a range of jobs. We considered Fleishman's taxonomy of 52 human abilities [4] which includes skills such as verbal comprehension and selective attention. Ten of these, which are motor skills such as finger dexterity and arm steadiness, constitute the most widely recognized taxonomy of skills. These ten skills also figure prominently in the O*NET job and skill database.

It was found in [6] that four of these ten motor skills consistently predicted job performance based on empirical evidence. The four skills reported to correlate consistently with job performance are - finger dexterity, manual dexterity, wrist finger speed and multiple coordination (see Table 1). Detailed definitions of these skills can be obtained in [4]. In brief, finger dexterity refers to the accuracy in finger movements while manual dexterity refers to the speed of arm movements. Wrist finger speed refers to the speed of wrist and finger movements and multiple coordination refers to the proficiency in performing coordinated movements with two or more limbs.

A large number of manual tests have been used to measure these motor skills. In all these tests, a candidate is asked to perform a task and is rated on the time taken to complete it and the accuracy achieved, if applicable. For example, one test to measure manual dexterity requires a candidate to unscrew pegs from one board, turn them over and attach them to another board [6]. A test for finger dexterity requires a candidate to insert a rivet in a hole and secure it with a washer, where this process is repeated multiple times. These tests measuring motor skills correlate with job performance in the range of $0.06 - 0.30$ (Table 1).

We seek to develop automated assessments to measure these four skills, which could serve as an alternate to the manual tests described. Our intuition is that these skills involve movements of different joints: wrist/finger accuracy and speed - movements of finger and wrist joints; manual dexterity - movement of shoulder and elbow joints and multiple coordination - coordinated manual dexterity. We develop apps based on this intuition. We limited our work to the action of hands and no other limbs.

Skill	Correlations [min-max]	Weighted Mean Correlations
Finger Dexterity	$0.07 - 0.21$	0.19
Manual Dexterity	$0.08 - 0.24$	0.22
Wrist-Finger Speed	$0.14 - 0.30$	0.18
Multiple Coordination	$0.06 - 0.15$	0.14

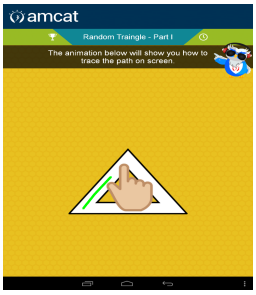
Table 1: Skills and their minimum, maximum and weighted average correlation values with job performance [6].

3. DESIGN OF APPS

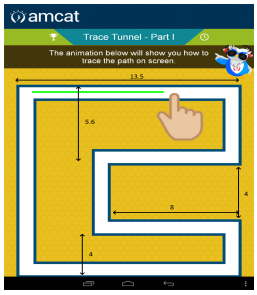
In this section, we describe the design of our touch screen apps to measure motor skills. We constructed each app to elicit specific hand and finger movements. We considered the simplicity and ease of comprehension of the apps as a key criterion. One should not be penalized for not understanding what has to be done, which could happen as a result of either cognitive or knowledge limitations. A set of instructions and a video/animation was shown before each app, to show how to perform the task. Each of these apps is described below:

1. **Douse the Fire (DOUSE)**: In this app, the candidate is shown 'fire' at random spots on a house shown on the screen (see Figure 1a). A candidate has to tap on the fire to douse it. As soon as the fire is doused at one spot, it appears at another spot on the house. In order to ensure that the fire occurs randomly, the distance between the two spots is probabilistically controlled using a uniform distribution between 0 and a number. The candidate has to douse as many fires within 30 seconds. We observed that the task requires elbow and shoulder movements and thus possibly measures manual dexterity.
2. **Trace a triangle-A (TRI_A)**: In this app, the candidate traces a path shown on the screen by dragging a finger over it. We initially considered having the candidate trace a line. However, we recognized that a candidate could not do this accurately because of the large surface area of the finger tip, restricting visual feedback of performing the activity incorrectly. We thus modified our exercise to contain two concentric equilateral triangles. The candidate was required to trace the path in between the triangles (see Figure 1b). The candidate was given feedback on the path traced by her through the use of colors. The path traced was green as long as it was confined to the designated area (space between the concentric triangles) and would turn red as soon as it went off the area. The width of the path is set to be more than the width of the fingertip (roughly 1 cm) to keep the task simple. The edge-lengths of the inner and outer triangles were 4.2 cm and 5.8 cm respectively. The candidate has

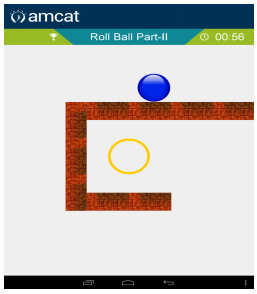
(a) Douse the Fire



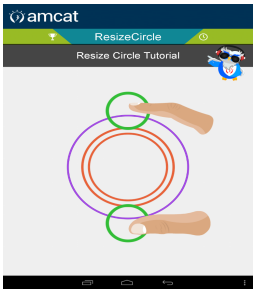
(b) Trace Triangle-A



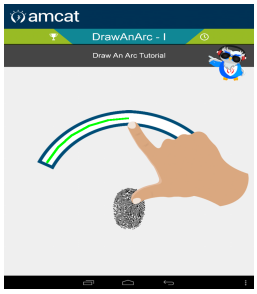
(c) Trace Path-A. (All dimensions are in cm)



(d) Roll Ball-B



(e) Resize



(f) Draw an arc-A

Figure 1: Snapshots of the apps

to trace as many triangles as possible in 30 seconds. As soon as one triangle was traced, another would appear. The app required moving one's hand quickly to trace the triangles and was designed to measure the speed element of manual dexterity. In principle, the task could be completed by finger movements, but we found that the default action made by the candidates which was comfortable to them involved shoulder and elbow movements.

3. **Trace a triangle-B (TRI_B)**: This app is similar to TRI_A with a difference that the width of the path was decreased. The width was kept a little lesser than the width of the finger tip. We hypothesized that the app required *careful* tracing and measured the accuracy element of manual dexterity.

4. **Trace a path-A and B (PATH_A and PATH_B)**: These apps are similar to the previous triangle apps. The difference is that candidates would trace over paths of much larger concentric polygons instead of a triangle, which shall require arm/hand movements (see Figure 1c). The polygons included rectangles, ellipses and those having zig-zag patterns. Figure 1c describes the dimensions of a sample path which was used. The path width shown in PATH_A is larger than those shown in PATH_B. The candidate has a maximum of two minutes to complete both the exercises and is required to trace as many polygons in the least possible time. These apps are designed to measure manual dexterity by tracing larger lengths and shapes, requiring different kinds of manual movements.

5. **Roll the ball-A (ROLL_A)**: In this app, a circle (symbolizing a hole) is positioned at the center of the screen and a ball is positioned at one of its corner. The ball rolls around on the screen on tilting the surface of the tablet. This is based on the tablet's accelerometer readings.[2] The candidate is required to guide the ball completely inside the circle. On the completion of one such exercise, the screen is refreshed with the ball placed at another point on the screen. The candidate has to complete four such exercises in the least possible time. The total time allotted is 40 seconds. The candidate moves the tablet with both her hands to guide the ball in the right direction. The test hence measures multiple coordination.

6. **Roll the ball-B (ROLL_B)**: This app is similar to ROLL_A. In this app, obstructions are placed in the path of the ball's movement (see Figure 1d). This is introduced to increase the degree of difficulty of the exercise. The time allotted to complete this exercise is 60 seconds.

7. **Fit a circle (FIT)**: We designed an app similar to the act of grabbing an object. The candidate is asked to perform a pinching action in a controlled environment. Two concentric circles were shown on the screen. The diameter of the inner concentric circle was fixed while that of the outer circle could be changed by the candidate. In order to change the diameter, the candidate had to place her thumb and her index finger on two points provided on its circumference and move them inwards or outwards without lifting them up. The diameter changed as the person dragged the two points.

[2]The accelerometer is calibrated at the beginning of the test by asking the candidate to place it on a flat table.

Skill Type	TT
Spot	Douse the Fire
Trace	Trace Triangle A and B
	Trace Path A and B
Multiple	Roll Ball - A and B
Grab/Pinch	Fit Circle
	Resize Circle
Rotate	Draw an Arc - A
	Draw an Arc - B

Table 2: List of tablet-based tests (TT).

The objective was to reduce the outer circle's circumference to match that of the inner circle. As soon as the two circles coincide, the screen is refreshed with two circles of different radii picked randomly. The candidate was required to perform this pinching action as many times as possible in 40 seconds. The app requires the rapid movements of fingers, say in grabbing many objects, one after the other and thus measures wrist-finger speed.

8. **Resize the circle (RESIZE)**: This is similar to the FIT app. The difference is that the outer circle now has to be shrunk and fit into a target ring as against placing it in a smaller concentric circle (see Figure 1e). On placing the outer circle within the target ring, the candidate is expected to lift her fingers from the screen, which then triggers the appearance of another target ring on the screen. The action is not considered until the fingers are lifted from the screen. This test measures the accuracy aspect, i.e. finger dexterity.

9. **Draw an arc-A (ARC_A)**: This app attempts to capture a candidate's wrist and finger rotation movement, as required, say, to screw or unscrew a nut and bolt. An arc is shown on the screen along with a pivot point (see Figure 1f). The candidate has to place her thumb on the pivot point and trace an arc shaped path with her index finger. On completing a trace, the screen is refreshed and a path with a different radius is presented. The candidate is required to trace six paths of varying radii in the least possible time. The arc paths are narrow (0.8 cm) requiring the candidate to be precise in her tracing. The entire task needs to be completed within 30 seconds. This test requires controlled and precise circular movements of the fingers. This test measures finger dexterity.

10. **Draw an arc-B (ARC_B)**: This app is similar to the ARC_A app but has wider arc paths. These arcs have 200% wider paths as compared to the arc paths presented in ARC_A. The candidate is required to trace as many arcs as possible in 30 seconds. This test requires rapid movement of wrists, say, in screwing a light bulb into a socket. This test measures wrist finger speed.

For each app, the candidate is instructed whether to place the tab on a table or hold it in her hands.

Skills measured	MST
Finger dexterity	O'Connor Tweezer Dexterity test [12]
Manual dexterity	GATB Manual Dexterity test [6]
Wrist-finger speed	Large Tapping test [6]
Multiple coordination	Purdue Pegboard test [13] We used the specific part of the test corresponding to coordination of both hands.

Table 3: List of non-automated manual motor skill tests (MST).

#	App	Score
1	Douse the Fire	Number of Correct douses
2	Trace Triangle - A	In-distance - Out-distance
3	Trace Triangle - B	In-distance
4	Trace Path - A	$\frac{\text{Time}}{\text{In-distance}-\text{Out-distance}}$
5	Trace Path - B	$\frac{\text{Time}}{\text{In-distance}}$
6	Roll Ball - A	$\frac{\text{Number of Rolls}}{\text{Time taken}}$
7	Roll Ball - B	$\frac{\text{Number of Rolls}}{\text{Time taken}}$
8	Fit Circle	$\frac{1}{\text{Number of fits}}$
9	Resize Circle	$\frac{1}{\text{Number of resizes}}$
10	Draw an Arc - A	$\frac{\text{Arcs}}{\text{Time taken}}$
11	Draw an Arc - B	In-distance

Table 4: Selected scores for each app. In-distance: Distance traced within path. Out-distance: Distance traced outside path.

4. EXPERIMENTS

We wish to answer whether the performance on tablet-based tasks can predict job performance. Specifically, we find out how our tablet-based tests and manual, non-automated motor skill tests compare in predicting job performance in industrial tasks like operating a lathe machine or tailoring clothes. This would act as a true indictor to suggest the practical use of the tablet-based tests in talent hiring. We note here that critical steps of non-automated motor skill tests like setting up the equipment, conducting the exercises and reporting scores are prone to human errors. Tablet-based tests have the distinct advantage of being devoid of such standardization issues. This advantage is likely to contribute towards its better predictive power.

4.1 Setup

The tests were administered to a workforce (referred to as *candidates* henceforth) belonging to three different occupations - tailors at a garment manufacturer, machinists and grinders at a machine-shop training company and machine operators at a skill training company. Each candidate was administered two sets of tests - tablet-based tests (TT henceforth) and non-automated, manual motor skill tests (MST henceforth). Four tests, as described in Table 3, were part of the MSTs. The standard set-up as described in [6] was followed in administering these tests. The eleven app-based tests described in §3 were part of the TTs.

TT and MST scores: In order to quantify a candidate's performance on our apps, we derived a single score for each

Job Performance Metrics	TT Scores				MST Scores				ATD Scores	
	Spot	Trace	Grab/Pinch	Rotate	MD	WFS	FD	MC	ATD	ATT
Tailors ($N = 74$)(Age range: $20 - 55$ years)										
Rate the tailor on the neatness of his/her completed work.	**0.22***	**0.37****	0.08	0.08	-0.09	-0.09	0.10	-0.10	NA	NA
Would you entrust him/her with a complicated task?	0.16	**0.33****	-0.10	-0.04	-0.08	**-0.33****	0.20*	-0.14	NA	NA
Rate how quickly s/he is able to complete her/his tasks.	**0.21***	**0.21***	-0.02	0.04	-0.13	**-0.20***	0.01	-0.13	NA	NA
Machinists and Grinders ($N = 68$)(Age range: $17 - 24$ years)										
Practical scores	**0.38****	**0.29****	**0.34****	0.07	0.07	-0.14	0.13	0.02	-0.06	NA
Electric Machine Shop score	**0.27****	0.11	**0.21***	-0.15	0.22*	**-0.29****	-0.03	0.10	0.12	NA
Machine Operators ($N = 78$)(Age range: $19 - 38$ years)										
Is s/he able to finish all the subtasks in a given operation?	0.15	**0.23****	0.00	-0.02	0.05	0.00	0.11	0.01	**0.20***	**0.27****
Rate how quickly s/he is able to complete the assigned operations.	0.17	**0.19***	0.04	-0.07	-0.14	**-0.19***	-0.01	-0.03	0.06	NA

* p<0.1; ** p<0.05; ATD : Attention to Detail scores; ATT : ATD + Best TT score;

FD - Finger Dexterity; WFS - Wrist-Finger Speed; MD - Manual Dexterity; MC - Manual Coordination.

Table 5: Correlations with job performance.

app (tabulated in Table 4). Further, the 11 tests were grouped into 5 skill types: *Spot* (DOUSE), *Trace* (TRI_A, TRI_B, PATH_A, PATH_B), *Multiple* (ROLL_A, ROLL_B), *Grab/Pinch* (FIT, RESIZE) and *Rotate* (ARC_A, ARC_B) (see Table 2). Each of these 5 skills was represented by a separate score. These scores were calculated by averaging the z-scores of apps contained in the skill[3]. For the four MSTs, scores were calculated as described in [6]. They generally measured the time taken to complete the task.

4.2 Data Set

The tests were administered to candidates belonging to three different occupations - 81 tailors, 74 machinists and grinders and 82 machine operators. The sample size was limited by the strength of the organizations. All three tests were administered by two event managers who had received a week's training on setting up the tests. Candidates performed the two tests (TTs and MSTs) with a gap of 5-6 hours. Each candidate's test was fully video-recorded. A review of these videos revealed that the standard process was not followed in 7.2% of the sample. These were discarded. The time recorded in nearly 3.7% samples for one or more of the MSTs was corrected. Post these changes, we finally had samples from 74 tailors, 68 machinists and grinders and 78 machine operators. We only considered the dominant hand in our analysis, except in the case for *multiple-coordination* which involves co-ordination between both hands. For machinists, grinders and machine operators, we also administered a multiple choice test of attention to detail (ATD)[4]. This was done to find what additional predictive power the TT scores added over the cognitive ability test scores to predict job performance.

Job performance scores: In the case of tailors and machine operators, a performance questionnaire (column 1 of Table 5) was developed on discussing with the candidates' managers. The managers were then asked to score the candidates on these metrics on a scale of 1 to 5. In the case of machinists and grinders, the training organization had documented scores from the candidates' lab-sessions. These scores were based on their performance on various job tasks given to them during their training. These ratings and scores formed the job performance data for our analysis.

4.3 Analysis and Observations

We compute the Pearson correlation coefficient (r) of all TT scores, MST scores and ATD scores (where available) with each metric contributing to job performance. The TT scores are fashioned to signal higher skill with higher magnitude whereas MST scores are fashioned to signal lower skill with higher magnitude. Hence, the correlation of job performance scores with TT scores is expected to be positive while the correlation with MST scores is expected to be negative. In our analysis, we observed the correlation between TT scores and MST scores to be in the range -0.27 to -0.34. This shows shared variance between the two scores. We noticed however that the scores of Multiple Coordination (one of the TTs) correlated positively with other MST scores. We hence do not include it in any further analysis. Additionally, by doing a regression, we found what incremental value the best correlating TT scores added over and above the ATD scores. These values and their respective significances are reported in Table 5.

First, and most importantly, we find that for every job per-

[3]Considering scores separately added no insight but increased complexity

[4]This is a criterion valid test used in hiring professionals in retail, sales, marketing etc. The 74 tailors had no formal education and hence could not take this test.

formance metric, at least one TT score shows a significant correlation (at $p \leq 0.1$) ranging from $0.19 - 0.38$ (mean: 0.27). This clearly establishes that TT scores are able to predict job performance and can be used for hiring/selection decisions by following standard practices. Second, MST scores show a significant correlation with four out of the seven performance metrics, where they range from -0.19 to -0.33 (mean:-0.25). We note here that the correlations between the four MSTs and job performance scores are in line with historically observed values (Table 1). ATD scores show a significant correlation in one case, where the *Trace* score adds significant incremental correlation (0.07) over and above it (column ATT, Table 5).

Among the app scores, there is maximum support for the *Trace* app which shows the highest correlation with job performance in five out of the seven metrics. In the remaining two metrics, the *Spot* app scores show the maximum correlation with job performance. While there is some support for the *Grab/Pinch* scores, there is hardly any support for the *Rotate* app scores. Among MST scores, the *Wrist-finger Speed* scores consistently correlate with job performance.

Discussion: We find that the TT scores are predictive of job performance in all cases in our study. The validity indices are comparable (and in cases bests) those observed for MST scores in the past (Table 1). The maximum support is for the *Trace* app. These are extremely encouraging results. This implies that the test may practically be used in making hiring decisions. The best way to do this would be to first perform a validity study with incumbents in a job in order to establish which TT apps distinguish on-job performance. These apps could then be used on new applicants and their scores be considered in the hiring process. While there is evidence for the *Trace* scores to be a universal predictor, the same may be established with further validity studies and meta-analysis. We envision that through such extended studies, a mapping could be formed between job roles and TT scores, akin to what has been established for MST scores. One would then know a priori which TT app and scores to use when hiring for a particular job role.

In four out of seven metrics, the MST and TT scores do equally well. One may observe that the MST scores did not do as well as the TT scores in three cases. This was surprising to us. A couple of reasons could explain this - first, the TT scores measure a larger variety of movements than the MST scores and some of these could potentially correlate better with job performance. For instance, there isn't any MST task similar to the structured tracing task in the TT. The other reason, as noted earlier, could be non-standardization and human errors in MST as compared to a controlled, completely standardized tablet-based test.

5. CONCLUSION AND FUTURE WORK

In this work, we explore the use of touch screen surfaces to measure motor skills. We show the scores of blue-collar workers on tasks performed on touch screen tablets to correlate with their respective job performances in the range of 0.19 to 0.38. These results make a strong case for using such automated, touch-screen based tests in job selection processes and in providing automated feedback. Such tests would make the process of identifying and credentialing skilled labor highly scalable and efficient, thereby benefiting both, individuals and corporations.

Our current work paves the way for substantial future work. The design of novel apps for motor skill measurement is a nascent area of research and could be further developed. By analyzing scores from such apps, we could create a map to suggest what scores are suitable for a given job role. Having such a map would help in automatically providing feedback to candidates on the skills they have. We could also perform the current tab tests for a number of other different job roles, which would help validate its design. Other devices and technologies such as smartphones[5] and resistive touchscreens could be experimented with, which could potentially make these tests more accessible, help do more accurate assessment and also grade new skills. For instance, a pressure detecting screen may help measure how soft the touch is, which might be relevant in nursing. We believe that the ideas introduced in this work can lead to substantial innovations in the blue-collar labor market.

6. ACKNOWLEDGEMENT

We thank Shashank Srikant for assistance and comments that greatly improved the manuscript.

7. REFERENCES

[1] J. Briel, K. OŠNeill, and J. Scheuneman. Gre technical manual. *Princeton, NJ: Educational Testing Service*, 1993.
[2] N. Claassen, M. De Beer, H. Hugo, and H. Meyer. Manual for the general scholastic aptitude test. *Pretoria: Human Sciences Research Council*, 1998.
[3] R. D. et al. The world at work: Jobs, pay, and skills for 3.5 billion people, 2012.
[4] I. Industrial/Organizational Solutions. Fleishman's taxonomy of human abilities, 2010.
[5] S. B. Issenberg, W. C. McGaghie, I. R. Hart, J. W. Mayer, J. M. Felner, E. R. Petrusa, R. A. Waugh, D. D. Brown, R. R. Safford, I. H. Gessner, et al. Simulation technology for health care professional skills training and assessment. *Jama*, 282(9):861–866, 1999.
[6] J. J. McHenry and S. R. Rose. Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification. Technical report, DTIC Document, 1988.
[7] S. L. McPherson and K. E. French. Changes in cognitive strategies and motor skill in tennis. *Journal of Sport & Exercise Psychology*, 13(1), 1991.
[8] A. Minds. Amcat. `https://www.aspiringminds.com/`.
[9] A. Minds. Skills plumbers 2015 report, 2015. `http://www.aspiringminds.com/research-reports`.
[10] K. Mononen. *The effects of augmented feedback on motor skill learning in shooting: A feedback training intervention among inexperienced rifle shooters*. University of Jyväskylä, 2007.
[11] O. I. Network. O*net online, 1998. `https://www.onetonline.org`.
[12] J. O'Connor. Instructions for the o'connor tweezer dexterity test. *Indiana: Lafayette Instrument*, pages 1–5, 1998.
[13] J. Tiffin. *Purdue pegboard examiner manual*. Science Research Associates, 1968.
[14] L. G. Ungerleider. Functional mri evidence for adult motor cortex plasticity during motor skill learning. *Nature*, 377(6545):155–158, 1995.

[5]Most apps here can be used on a smartphone with some adjustment in the scale and aspect ratio of the apps and recalibration of scores. It may not effectively measure wider movements of the arms.

Industry Track - Posters

Studying Assignment Size and Student Performance Using Propensity Score Matching

Shirin Mojarad
McGraw-Hill Education
281 Summer Street,
Boston, MA USA
Shirin.mojarad@mheducation.com

ABSTRACT

Teachers and instructors assign students homework of varying lengths. There is considerable evidence that factors such as cognitive load play a role in student performance and learning, but there has not been sufficient study of how these phenomena play out in the specific case of the length of homework. In this paper, we study the impact of assignment size on student performance. This paper represents the first attempt we are aware of to study how long assignments should be, in real-world data, in order to maximize student performance and learning. However, natural assignments of different lengths often vary in other ways. We control for this limitation using propensity score matching (PSM), an approach that helps to control for variables affecting outcome besides the intervention of interest. As such, we can conduct our analysis on large-scale data naturalistically collected through a digital educational platform. We use PSM to study the effect of assignment size on student performance while controlling for assignment difficulty, discrimination and reliability. We find that shorter assignments result in higher performance. These results can be used as a guideline for instructors and instructional designers when designing course assignments.

Keywords

Propensity score matching, assignment size, classical item analysis, item difficulty, item discrimination, student performance, test reliability

1. INTRODUCTION

Graded assignments are used as an effective method to improve students' performance on final tests and improve learning [1]. Considering multiple shorter assignments as opposed to few, larger assignments is amongst the recommendations by USC for designing effective homework assignments [2]. This is because shorter assignments are less intimidating and help enhancing student motivation by minimizing the negative effects of a poor grade on student learning experience. In this study we investigate whether assignment size affects student performance. Since assignments of different sizes often vary in other ways, other assignment characteristics affecting the performance should be isolated to enable the study of assignment size effect on student performance.

Randomized control trials are considered the gold standard in conducting studies to investigate the effect of a particular intervention on a specific outcome [3]. However, their application is limited in educational settings as they can be conducted on a limited number of students. Results from the comparison of RCTs and OSs show that OSs can expand upon RCTs due to the use of large and diverse sample population [4]. Propensity score matching (PSM) is a common method in OSs to study the causal effect of an intervention on a particular outcome [4]. In this paper, we have used PSM to leverage the large amounts of data available through McGraw Hill Education digital platforms.

The goal of this paper is to study the impact of assignment size on student performance in isolation from other assignment characteristics including assignment difficulty, discrimination and reliability. This is the first effort of its kind in measuring an optimal assignment size to maximize student performance.

2. Materials and Methods

2.1 Data

We study these issues using data from assignments completed through McGraw-Hill Education's higher education platform, Connect. Connect is one of the most widely used digital platforms in higher education with over two million students and 25,000 instructors [5][6]. Connect allows instructors to design assignments in form of homework, practice, exams, or quizzes. Here, we refer to assignments as a set of items that either test student on knowledge and skills or allow students to practice what they have learnt on the course. Most Connect assignments are graded by the system automatically.

The dataset in this study is retrieved from all the courses created from the title Managerial Accounting 2nd Edition, by Robert Libby. We include all data for this title between September 2014 and January 2016. The original dataset included 362 classes, where 12,588 students responded on 3,072 items on 5,330 assignments, for a total of 1,031,298 student-item pairs. We have kept only assignments that have 10 or more student submissions. After applying this filter, there are 2,826 assignments left in the data. From the four of types of assignments in Connect, i.e. homework, practice, exam and quiz, we have focused on homework assignments. The reason is that in homework assignments, score is not as a strong motivator as exams and quizzes since homework assignments have a low weight in final score and are mainly aimed for development of self-study habits in students [7]. Hence, students are more motivated by learning to finish homework assignments. Therefore, size of a homework assignment will be an important factor in keeping students engaged throughout the assignment.

2.2 Exploratory Data Analysis

The dataset includes assignments' size, difficulty, discrimination, reliability, and average score where difficulty, discrimination and reliability are calculated using classical item analysis [8]. The assignment size in this dataset varies between 1 to 101 items.

Based on the rarity of very large assignments (and the likelihood that an assignment with over 100 items represents test practice or something different than briefer assignments), we have filtered down to assignments of size 16 or less. Filtering in this fashion still retains 98% of the assignments. We categorized assignments into short and long assignments by using a cut off for number of items within that assignment. Frequency of assignments drops for assignment sizes of larger than 5, which indicates most instructors prefer shorter assignments of size 5 or less. We have used this as a reference to decide a cut off value for number of items for short and long assignments. Following this definition, there were 1,787 short and 1,039 long assignments.

Figure 4 shows the mean score of different assignment sizes. As shown in this figure, the mean score of assignments drops as the assignment size increases.

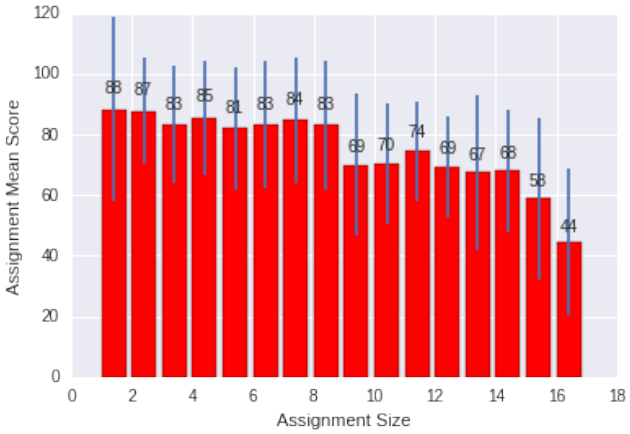


Figure 4. Assignment size versus assignment mean score

3. Results

Overall, students achieve an average 8.7 (on a scale of 0 to 100) higher score on short assignments than long assignments. When we control for difficulty, discrimination and reliability using PSM, students still achieve a 6.8 (on a scale of 0 to 100) higher average score on short assignments compared to long assignments.

The differences between the characteristics of short and long assignments matched using PSM are shown in Table 2. We have used Algina's d to compute the effect size of the difference of means between the two assignment groups [9].

As shown in this table, the effect size of difficulty, discrimination and reliability between two groups of assignments is negligible, indicating that these factors are no longer significant once we control for them using propensity score matching.

Table 2. P-value and the effect size of short versus long assignments, matched using PSM method

Attribute	Mean Difference	Effect Size (Algina's d)	P-value
Average Score	6.8	0.40	<0.001
Difficulty	0.00	0.00	0.99
Discrimination	0.00	0.01	0.53
Reliability	0.00	- 0.01	0.44

4. Conclusion

In this study, we investigated the effect of assignment size on student performance. Results of EDA show that student performance drops as the assignment size increases. The relation between assignment size and average score indicated that performance drops dramatically in assignments sizes of higher than 6. Hence, we used a cut off value of 6 to define short and long assignments. In order to investigate the statistical significance of this difference in two groups of assignments, in isolation from other factors affecting assignment performance, we used propensity score matching (PSM). The effect size and average performance difference of short versus long assignments is still significant when matching assignments with similar difficulty, discrimination and reliability. This indicates that longer assignments may increase cognitive load for students and negatively affect student performance and learning. These results can be used in form of recommendations to instructors when they are designing homework assignments on the Connect platform.

5. ACKNOWLEDGMENTS

This research paper is made possible through the help and support from Professor Ryan Baker, Dr. Lalitha Agnihotri and Alfred Essa, VP Analytics and R&D at McGraw-Hill Education.

This paper is based on work supported by the McGraw-Hill Education Digital Platform Group. Despite provided support, any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect positions or policies of the company.

6. REFERENCES

[1] Latif, E., Miles, S., (2011). The impact of assignments on academic performance. *Journal of Economic and Economic Education Research*, (Nov. 2011), 12 (3).

[2] Center for Excellence and Teaching, University of Southern California,http://cet.usc.edu/resources/teaching_learning/docs/teaching_nuggets_docs/4.2_Assignments_and_Homework.pdf, last accessed at March 2016.

[3] Silverman, S. L. (2009), From Randomized Controlled Trials to Observational Studies, *The American Journal of Medicine*, Volume 122, Issue 2, Pages 114–120.

[4] Rosenbaum, P. R., and Rubin, D. B., (1983). The Central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41-55.

[5] Feild, J., 2015. Improving student performance using nudges analytics. *Educational Data Mining*.

[6] Agnihotri, L., Aghababyan, A., Mojarad, S., Riedesel, M. and Essa, A., (2015). Mining Login Data For Actionable Student Insight. In Proc. 8th International Conference on Educational Data Mining.

[7] Sharma, Y. K., Fundamental Aspects of Educational Technology. Kanishka Publishers, Distributors New Delhi.

[8] Smith, Jeffrey K.. (1987). Review of *Introduction to Classical and Modern Test Theory*. *Journal of Educational Measurement* 24 (4). [National Council on Measurement in Education, Wiley]: 371–74.

[9] Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317–328.

Toward Automated Support for Teacher-Facilitated Formative Feedback on Student Writing

Jennifer Sabourin
SAS Institute
100 SAS Campus Dr.
Cary, NC 27513
1.919.531.3313
Jennifer.Sabourin@sas.com

Lucy Kosturko
SAS Institute
100 SAS Campus Dr.
Cary, NC 27513
1.919.531.3430
Lucy.Kosturko@sas.com

Kristin Hoffmann
NC State University
Raleigh, NC 27695
1.919.515.7061
klhoffma@ncsu.edu

Scott McQuiggan
SAS Institute
100 SAS Campus Dr.
Cary, NC 27513
1.919.531.1119
Scott.McQuiggan@sas.com

ABSTRACT

Formative, content-level feedback on student writing has been shown to have positive impacts on both writing and learning outcomes. However, many teachers struggle to provide this type of feedback to large classrooms of students. This paper takes an initial step towards supporting teacher-facilitated feedback through the use of automated and user-directed topic discovery. 114 student essays were collected from a local underperforming middle school as part of a pilot study for Write Local, a digital repository and workspace for authentic problem-based learning activities. Predictive models were built and evaluated to explore the impact of different topic discovery approaches as well as correction of student spelling errors on model accuracy. The resulting models provide promising direction for scaffolding teachers in providing formative feedback on content-level features of students' problem-based writing.

Keywords

Problem-based writing, formative feedback, teacher-facilitated feedback, automated writing assessment, topic discovery

1. INTRODUCTION

Problem-based writing tasks seek to elicit high-quality student writing by contextualizing the purpose of the task and providing an authentic audience [4]. These tasks also tend to extend across several days or learning periods offering more opportunities for formative assessment and feedback, which is expected to yield improved writing outcomes [1]. However, it is often difficult for teachers to focus on high-level features such as the focus, accuracy, and organization of student writing when working with a large classroom of students. Instead, teachers are more likely to focus on surface level features such as spelling, grammar, and mechanics. This is especially true in underperforming schools [2].

This work serves as an initial investigation into automated assessment of student writing in order to scaffold teachers in providing higher-level formative feedback. A pilot study was conducted as an initial step in the Write Local project. Write Local is intended to be a digital repository and workspace to facilitate both teachers and students in authentic problem-based writing activities. As part of a pilot study, 114 student writing samples were collected from students at an underperforming [3], local middle school as part of a multi-day problem-based learning activity. Student essays were manually coded for essay focus and accuracy. A variety of models for predicting these features were constructed and evaluated as an initial exploration for scaffolding teacher-facilitated feedback. In particular, this work sought to explore the role of automated and user-directed topic discovery in predicting content-level essay features. Additionally, we sought to investigate the importance of correction of student spelling mistakes prior to model construction. The results indicate that these initial models can serve as a starting point for supporting teachers in providing feedback on content-level features in problem-based writing and inform several directions for future work.

2. PILOT STUDY

This investigation uses data collected during a pilot study of Write Local. Write Local seeks to employ crowdsourcing to ensure teachers and students have immediate access to a large repository of writing prompts that cover the entire spectrum of text types and audiences—persuasive, informative/explanatory and narrative. Local businesses, and in particular, those employing STEM-related positions, can post various letters of need as well as any supplemental documentation such as images or vocabulary lists. Teachers can then select a call from the repository and assign the project to their students. Students will use the integrated workspace to plan, research, document, draft, revise, present, and submit their response in one central space.

The entire sixth grade from a local, underperforming [3] middle school (54% free/reduced lunch) participated in this study as part of their regular social studies class. Of the 168 participants, 86 were male and 82 were female with a mean age of 11.5. Of the 168 participants, 114 completed all components of the procedure. For the remaining analyses only data from these 114 students is used.

For the study, students were divided by class into one of two conditions: experimental and control. On the first day of the study, students in the experimental condition viewed a 3-minute introduction video that contained problem context: a frozen yogurt company plans to open a new location and asked students to write a letter with their researched opinions about 1) which 5 toppings should be available on the topping bar and 2) where the new shop should be located. Students used authentic data and a map of the area to make their decisions. Students in the control condition were given a similar task without real-world contextualization. Students in both conditions were given two full 50-minute class periods to plan and write their letters.

Three researchers then transcribed and coded the essays with sufficient inter-rater reliability ($k = .89$). Essays were given a composite score for essay focus and accuracy. Using the final composite scores, students were divided into 3 evenly distributed categories (High, Medium, and Low) for both focus and accuracy. These groupings are intended to be presented to teachers to inform formative feedback for their students.

3. TEXT ANALYSIS AND MODELING

The first step in building predictive models of student essay content classifications was to extract meaningful features from the student text. In total, the corpus for analysis included 114 student essays. The average length of the essays was 130.0 (SD = 91.4) words and 9.6 (SD = 7.6) sentences. The writing samples provided by the students were analyzed using SAS® Text Miner® and SAS® Enterprise Miner®.

For the purpose of this analysis we focused on the document topic analysis features of SAS Text Miner. The text topic procedure identifies terms that are strongly associated within the corpus. It also provides a strength of each topics' presence within the document. Topics can be automatically learned from the corpus or they can be provided or fine-tuned manually. Both approaches were used for this work. For automatic topic discovery, the limits were set at 25 multi-term topics. Manually-created topics were generated by highlighting terms in the text of the prompt and identifying whether each term applied to the problem context, the problem request, or the task instructions. In total 27 terms were identified; 8 context terms, 13 request terms, and 6 instruction terms. These terms were provided as user-created topics to the topic discovery procedure. In addition, up to 25 multi-term topics could be automatically generated; though because the engine tries to remove correlated topics, only 22 new topics were created. Of the 27 user-provided topics, only 17 occurred in the corpus of student data; 6 context terms, 9 request terms, and 2 instruction terms.

During essay transcription and coding, it was noted that there were a significant number of spelling errors present in the corpus. This may be due to the fact that essays were handwritten without the support of automated spell checking tools that many students are familiar with. In order to investigate the importance of correct spelling in modeling content-level features such as essay focus and accuracy, we chose to build models using different levels of spelling correction. Three different corpora of student essays were provided to the text topic discovery procedures: 1) the students' original texts, 2) an automatically spell-corrected version of the text, and 3) a manually spell-corrected version of the text.

For this exploration, we evaluated models across both topic discovery type (fully-automated and user-facilitated) and spelling correction type (manual, automated, and no correction). Additionally, we built separate models to predict both essay focus classification and essay accuracy classification. Finally, we used three modeling approaches for each corpus: logistic regression, decision tree, and neural network.

Each model was evaluated using 10-fold cross validation and predictive accuracies were compared against a baseline of most frequent class. This measure was 33.0% and 40.4% for essay focus and accuracy, respectively. The most common class for each evaluation type was Medium. With one exception, all models outperformed baseline with statistical significance at the 0.05 level (Table 1).

Overall, the models built using manual spelling correction and prompt-based topics outperformed other models in predicting essay focus and accuracy. This suggests that the prompt-based topics centered on the components of problem-based learning activities were beneficial in improving predictive accuracy. Unfortunately, this step requires manual annotation for each prompt. At present, this task, while manual, is not particularly labor intensive and can scale as we assess whether this benefit holds for future, unseen prompts. However, since the objective of Write Local is to scale with a large number of problem-based prompts, this may no longer be feasible. If we determine that this type of prompt annotation continues to be beneficial for predicting essay accuracy and focus we may investigate possible methods for automating or facilitating this task.

Table 1. Predictive accuracy for essay focus and accuracy using (a) discovered topics and (b) prompt-based topics

Discovered Topics

	Spelling Correction		
Model	Manual	Auto	None
Neural Net	F: 57.4 A: 55.0	F: 48.9 A: 51.9	F: 45.5 A: 52.6
Log. Reg.	F: 46.5 A: 56.1	F: 45.5 A: 46.4	F: 44.6 A: 47.3
Decision Tree	F: 51.3 A: 55.2	F: 46.4 A: 45.3	F: 47.3 A: 43.9
Average	F: 48.9 A: 55.7	F: 46.9 A: 47.9	F: 45.8 A: 47.9

Prompt-Based Topics

	Spelling Correction		
Model	Manual	Auto	None
Neural Net	F: 61.4 A: 56.1	F: 50.8 A: 57.1	F: 55.4 A: 50.0
Log. Reg.	F: 56.1 A: 68.4	F: 46.4 A: 62.5	F: 49.1 A: 52.7
Decision Tree	F: 53.5 A: 61.4	F: 50.0 A: 54.5	F: 46.5 A: 57.1
Average	F: 57.0 A: 62.0	F: 49.1 A: 58.0	F: 50.3 A: 53.3

Secondly, we note that the models using manual spelling correction tended to outperform models using automatic or no spelling correction, though this finding was less reliable. Since the "manual" spelling correction was done primarily using feedback from a word processor, it may be the case that had the essays been written digitally with spell check options available, many of the errors that were corrected would have been found by the student themselves. Future work will be necessary to determine if word processor spell check features are sufficient for this task.

4. REFERENCES

[1] Graham, S. et al. 2015. Formative Assessment and Writing: A Meta-Analysis. *The Elementary School Journal*. 115, 4 (2015), 523–547.

[2] Matsumura, L.C. et al. 2002. Teacher Feedback, Writing Assignment Quality, and Third-Grade Students' Revision in Lower-And Higher-Achieving Urban Schools. *The Elementary School Journal*. 103, 1 (2002), 3.

[3] North Carolina School Report Cards: *http://www.ncpublicschools.org/*. Accessed: 2016-03-01.

[4] Purcell-Gates, V. et al. 2007. Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly*. 42, 1 (2007), 8–45.

TutorSpace: Content-centric Platform for Enabling Blended Learning in Developing Countries

Kuldeep Yadav, Kundan Shrivastava, Ranjeet Kumar, Saurabh Srivastava, Om Deshmukh
Xerox Research Centre, India
kuldeep.r@xerox.com, om.deshmukh@xerox.com

ABSTRACT

One significant impact of the Massive Open Online Courses (MOOCs) phenomenon is that they have accelerated the widespread availability of quality education content. We refer to this content as the Open Educational Resources (OERs). It is our hypothesis that the OERs can be used to supplement classroom teaching for improved teacher efficiency and better student outcomes. We present a platform called TutorSpace which helps in curating OER content from multiple sources, integrating this content into a curricular setting in the context of what the lecturer is teaching and delivering it to students in a personalized way. A particular novelty of the TutorSpace platform is its capability for content-driven non-linear navigation of video content.

1. INTRODUCTION

The developing economies such as India, Brazil, China, etc face acute shortage of quality instructors, which is one of the primary reason for large number of unemployable graduates [2, 3]. Quality educational content (i.e. videos, slides, assignments) generated by the MOOCs can be potentially used to improve student learning and engagement in developing countries. However, instructors find it hard to use OER content directly in their course due to many reasons such as lack of context, no easy way of cross-source content aggregation, limited content search and curation capabilities of existing systems, and network bandwidth constraints. For example, $Alice$ is an instructor of an Algorithms course in XYZ university and she had taught some of the basic sorting algorithms to the students of her class. She wants to find specific videos for the "heap sort" algorithm concept, which can be given as an homework to the students. As, there would be different videos available online for this concept with varying duration, difficulty level, sources, etc. Alice is likely to spend a lot of time navigating through the available videos to finally select a video which suits her class' requirement.

We present a platform called *TutorSpace* that helps in searching and curating OER content from multiple sources, allows integration this content into a curricular setting in the context of what the lecturer is teaching and helps delivering it to students in a personalized way. TutorSpace uses advance multimedia concepts to support features such as quick and efficient video navigation, identification of topic transitions in a video, adding annotations on a video, etc. For the students, TutorSpace enables self-paced and ubiquitous learning where they can see course material posted by the instructor. TutorSpace also provides capabilities for students to share their notes, video bookmarks with their peers and discuss the topic of mutual interest in discussion forums.

2. TUTORSPACE PLATFORM

The proposed TutorSpace platform [1] provides content-centric capabilities to help instructors in the course curation. It allows instructors to have a digital presence of a classroom-based course, ability to search relevant course materials, and inclusion of selected education content in the curriculum. One of the key features of TutorSpace is that it provide a lecture planning workbench where the instructor can pool content from different sources and intersperse outside content with snippets of his/her pre-created content or classroom teaching. For students, TutorSpace enables self-paced and ubiquitous learning where they can see course material posted by the instructor. It also allows students to share their notes, video bookmarks with their peers and discuss topics of mutual interest in discussion forums. Some of the primary functional components of TutorSpace are as follows:

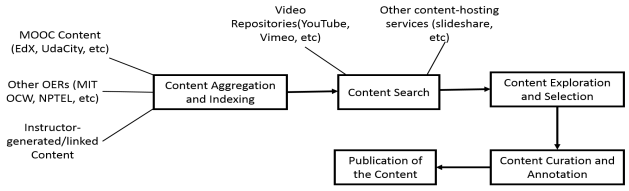
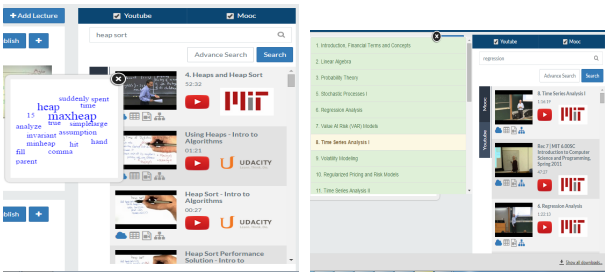


Figure 1: Step-by-step overview of instructor-led content curation and selection

2.1 Content Aggregation, Indexing & Search

TutorSpace aggregates content from different sources i.e. MOOCs (Coursera, EdX, Udacity), YouTube, etc. The content aggregation includes indexing meta-data about the course (i.e., information, syllabus), and video lecture specific meta-data (title, description, transcript of the video, duration, etc). Similarly, TutorSpace provides flexibility to the instructor to upload/link his/her own self-generated content too. Figure 2a presents a snapshot of the search dashboard in TutorSpace. Instructor can search for any concept and the system returns a set of relevant video lectures. The instructor has the flexibility to add search filters w.r.t. the source of the content (e.g., known-OER or all-YouTube) as well as other advanced filters such as duration, presentation style (e.g., slide or black-board), etc. Additionally, TutorSpace indexes meta-data about each video and further, this meta-data is presented to provide additional cues to the instructor as shown in Figure 2b. One of these cues is customized word-cloud which contain some of important concepts covered in the video (i.e., video preview). A detailed step-to-step creation process of customized word-cloud is presented in one of our earlier work [5]. These cues can help in the first-level decision making of whether to play a video or not. For example, word-cloud can help instructor in answering broad question about the video such as, "does this video contain algorithms for both linear and binary search" or "does this video explain heap sort with implementation



(a) (b)

Figure 2: (a) A snapshot of content search dashboard of TutorSpace. (b) Snapshot of concept relationship for a video

in Python programming language". In low bandwidth settings, it can save significant amount of time for the instructors [5].

2.2 Content Exploration and Selection

The instructors need to take a deep-dive and explore the content completely before including it in the teaching plan. Content exploration, specifically for a video, is a time-consuming task where often videos have long durations. The instructor can select any video for detailed exploration from the search results shown in Figure 2a. TutorSpace makes content exploration less time-consuming by providing techniques for non-linear navigation in a video with the help of customized word-cloud and parallel 2-D timeline as shown in Figure 3. Consider a video with the duration of nearly 60 minutes which discusses different sorting algorithms, the information provided by the customized word-cloud will include the name and time sequence of different algorithms along with other important terms discussed, which can help an instructor in getting a time-aware representation of a video [5]. Further, the customized word-cloud is interactive and instructor can click on any of keyword and its occurrences are highlighted on the 2-D timeline. The keyword occurrences represent different time instances where the keyword appears in the video. Further, mouse-hover event on any of these occurrences provide the context (i.e. an adjacent sentence) where a given keyword has been spoken. The click on any of occurrences will navigate the video to the point, where it was spoken in the video.

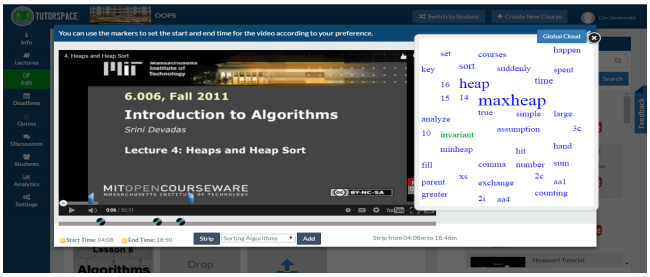


Figure 3: A snapshot of non-linear video navigation dashboard in TutorSpace with the help of customized word-cloud

Sometime, instructors may want to select a part of the content as opposed to the complete video. For example, in a 60 minute video on sorting algorithms, she may want to select only “merge sort" concept and share it with the students. TutorSpace enables partial selection of a content using its easy “video stripping" method. As shown in Figure 3, The instructor can move “start" and “end" (blue color) markers on the video timeline to highlight part of video content and click on “strip" button to select the content. After selecting the content, the instructor can drag and drop the content in their lecture plan as shown in Figure 4.

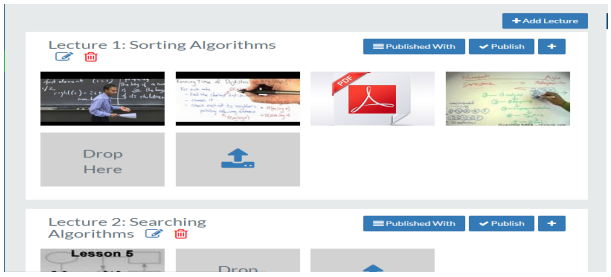


Figure 4: A snapshot of selected content (lecture plan) in TutorSpace

2.3 Other Features

TutorSpace provides a simple and user-friendly way to add notes and bookmarks on a video. After curation, the instructor can play the video and add annotations in terms of textual notes, images, external links/documents, etc with a click of a button. TutorSpace maintains detailed logs of interaction of the students with the content. It provides descriptive analytics on shared content to the instructors. The analytics include simple student-specific viewing statistics to fine-grained interaction pattern (i.e., time spent, pauses, play, etc). The instructor can use these findings to adapt the course curation strategies or to infer perceived difficulty of certain concepts. For example, if many students are spending a considerable amount of time on a specific portion of a video, it may need to be clarified during the class. Furthermore, TutorSpace provides standard learning management system (LMS) specific features such as course management, deadline creation and submission, quizzes, discussion forums, and student information management.

3. DISCUSSION

In developing countries such as India, quality of education is yet to improve substantially. We presented TutorSpace platform which can seamlessly enable integration of high-quality OER content in traditional classroom settings. TutorSpace provides rich multimedia capabilities w.r.t. content-indexing, search, non-linear navigation, and rich curation of the content. These capabilities are specifically designed to help instructor in developing countries. In our initial field-trial with the instructors, they appreciated the capabilities of the platform and provided several valuable feedback, which will be crucial for a long term acceptance of such a platform. We are in process of deploying TutorSpace to many engineering colleges in India and will be discussing our experiences in a future study.

4. REFERENCES

[1] TutorSpace project page, **http://xrci.xerox.com/tutorSpace-at-scale-personalized-learning**

[2] Cutrell, Edward et al. "Blended Learning in Indian Colleges with Massively Empowered Classroom." In Proceedings of the Second (2015) ACM Conference on Learning@ Scale, pp. 47-56. ACM, 2015.

[3] Chetlur, Malolan et al. "EduPaL: Enabling Blended Learning in Resource Constrained Environments." ACM DEV 2014.

[4] Guo, P. J., & Reinecke, K. (2014, March). Demographic differences in how students navigate through MOOCs. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 21-30). ACM. Chicago

[5] Yadav, K. et al. Content-driven Multi-modal Techniques for Non-linear Video Navigation. In Proceedings of the 20th International Conference on Intelligent User Interfaces (pp. 333-344). ACM.