Mining Large Scale Data from National Educational Achievement Tests

A Case Study

Reihaneh Rabbany, Osmar R. Zaïane
Department of Computing Science,
University of Alberta
{rabbanyk,zaiane}@ualberta.ca

Samira ElAtia*
Campus Saint Jean,
University of Alberta
selatia@ualberta.ca

ABSTRACT

Large scale analysis of educational assessment data, outlines patterns of success and failure, highlights factors of success, and enables performance prediction and eventually leads to proper ways of intervention. It has applications in both traditional settings where data is extracted from paper tests and surveys, and in e-learning settings such as distance, hybrid learning, and online courses. In the latter, drop out prediction and finding its factors and patterns is gaining much attention within the research community. In the former, the performance prediction is at the center of focus as drop outs are rare. Although the platform and data extraction is different, the essence of analyzing the test data is similar in both settings. In this paper, we present a case study on using data mining techniques in the analysis of large scale assessment data. The data is from the PanCanadian Assessment Program (PCAP), which is a national achievement tests administered by the Council of Ministers of Education, Canada (CMEC). The original findings published based on this data underwent rigorous traditional statistical analyses. Here, we show new insights that could be obtained from the same data, by leveraging the power of Data mining.

Keywords

Test Data, Large Scale Standardized Tests, Educational Assessment, Learning Analytic, Educational Data Mining

1. INTRODUCTION

Diaz and Brown [2012] define Learning Analytics (LA) as the "use of data, statistical analysis, and explanatory and predictive models to gain insights and act on complex issues (...) about the learners". Two types of data can be used for implementing LA in educational contexts. First, data generated by the learners themselves, and often referred to as digital footprints. This type of data would enable us to implement techniques to carry out data mining analyses leading to a holistic understanding of students' behaviour. Second, data supplied by learners in the form of surveys and other demographic and background information. This data provides a foundation for building an information system about the learners. Both types of data are necessary to learn about how learners react, behave, interact and use a specific e-learning environment. It also provides (1) insights on how effective are such environments and (2) feedback for both future improvement and potential wider use.

In this research project, we focus primarily on studying the potential implementation of LA to empirically study students performance on standardized educational achievement tests. To carry out the study, we are using the PCAP data from CMEC. The PCAP Data contains the two types of data mentioned above. First, we have students test results highlighting their performance on PCAP; and second, we have substantive background information from three surveys: students, teachers, and school administrators. The data collected is an indication of the performance of students on this national test, and provides a unique window into students performance in reading, mathematics and sciences in conjunction with qualitative input from other stakeholders.

With this paper, we demonstrate how data mining can be applied and used for analyzing test data combined with meta-data from the surveys. This case study presents how data mining facilitates learning analytics research and presents the potential of its use to advance research on assessment from a big picture perspective.

In the rest of this paper, we first overview the related works on educational data mining and learning analytics. Then we introduce the PCAP exams and briefly present the past published results on the PCAP data. With this background, we then move to present our case study, starting with the full description of the data, the data cleaning challenges, and then the analyses performed on the data. Main contribution of our paper could be summarized as:

- In general, providing various educational stakeholders and decision makers with new insights and connections in studying tests data
- For PCAP data, providing more accurate comparison of different factors by relaxing the assumptions about underlying population and using descriptive statistics
- Finding new insights about distribution of students, such as outliers students i.e. over/under achievers
- Ranking and grouping different factors that affect performance of students

*The corresponding author and project leader of the study.

ASSESS Data Mining for Educational Assessment and Feedback Workshop in conjunction with KDD 2014, New York City, August 24, 2014
.

2. RELATED WORKS

Han et al. [2006] define data mining as the "analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners." Likewise, Fayyad et al. [1996] emphasize the fact that the discovered knowledge that ensues from the data mining procedure has to be previously unknown, non-trivial, and genuinely useful to the data owners. Data mining techniques have a very broad range of applications: medical, biological, finance, industrial and corporate. *Educational applications* of data mining and learning analytics are on an emerging and growing trend due to the vast data becoming available from the growing number of courses delivered in e-learning environment. In addition, the new trend with MOOCs (Massive Open Online Courses) is accelerating the need for advanced and scalable data mining and analytics techniques for educational data.

2.1 EDM and Learning Analytics

Educational Data Mining (EDM) applies data mining techniques to explore data originating from educational contexts and to study educational questions [Romero and Ventura, 2010]. The unique hierarchical features of educational data [Baker, 2010] provide researchers in educational environment with opportunities to use data mining as a tool for investigation. EDM usually consists of implementing data mining techniques in case studies using sets of actual education data from an institution of higher education, in order to aid in decision making processes and improve the organizational effectiveness [Huebner, 2013].

The parallel field of Learning Analytics and Knowledge (LAK) focuses on collecting, measuring, and analyzing data about learners and their learning contexts for the purpose of optimizing these learning contexts. Learning Analytics is data analytics in the context of learning and education; that is, the collection of data about learners' activities and behaviour as well as data about the environment and context in which the learning took place; and the analysis of such data using statistics and data mining techniques with the purpose of extracting relevant patterns from this data to better understand the learning that took place. [Siemens and Baker, 2012] state that LAK bridges between the micro level of data mining and the macro level of educational research and aims to understand the learning environment.

The objectives of Learning Analytics can thus either be the reporting of measures and patterns from the data for the understanding of learning activities, or the optimization of the learning activities and strategies or the environments in which the learning occurs. For this project, we aim to target the first objectives of LAK.

2.2 DM and LA in Educational Assessment

We view Learning Analytics as an inquiry-based approach to research in educational setting. Dewey [1910] defined inquiry-based learning as a complex process whose aim is to find new meaning and connection and build knowledge and understanding. It is a process whereby the learners, students in this case, think about problems and situation, raise questions that they then try to solve and find answers to. Students are not limited to the objective of a restrictive curriculum, they learn from making connection between what is learned in the classroom and what they encounter in their daily lives. In light of this, LA can provide an inquiry-based approach to powerful learning experiences that would engage both educators and students. Merriam et al. [2012] assert that "Transformational Learning is about change, dramatic, fundamental change in the way we see ourselves and the world in which we live" (p. 123). Within an assessment for learning [Popham, 2008] framework, the two sources of the PCAP data provide a unique opportunity to transform a standardized achievement test to a formative assessment tool that allow various stakeholders (that include parents, teachers, administrators and decision makers) to not only revisit learning through assessment but to become involved in the assessment process [Popham, 2007].

For this study, we are transforming learning from assessment by looking at data from a different angle. While we are not trying to psychometrically analyze the students responses to tests and to questionnaire, we are studying the feasibility of DM algorithm to analyze this data in a transformative way. In measurement, various mathematical models are used to analyze the validity and the reliability of items in tests. Hambleton et al. [1991], Lord et al. [1968] provide a good references for such models. Hambleton and Jones [1993] provides excellent overview of Item Response Theory.

Gobert et al. [2013] developed an inquiry intelligent tutoring system to assess science inquiry skills. As assessment come to play a big role in educational setting, the approaches to analyzing assessment data would change. Pellegrino [2004] states that "I pose that it is not just a matter of quantity or quality. Rather, we can improve educational outcomes through assessment but only when we have better assessment practices.(p.5)" He listed three major elements that would cause a shift in handling data in assessment. As sciences and information technologies advance, "increases in computational power and statistical methods; the dynamics of population change, which will push us even greater in terms of the pursuit of equity and excellence; and, finally, the rhetoric and politics of accountability."

We are looking into large sets of data: one is quantitative data from examinees on test items, and the other is a qualitative data from the examinees themselves, their teachers, their school administrators and their parents. We did not seek the traditional psychometric analysis of the test results to check for validity and reliability [Linn and Gronlund, 2000], as this has been conducted by CMEC, but rather we wanted to exclusively investigate what can be revealed when both data sets are combined for Learning Analytics.

3. PANCANADIAN PROGRAM

PCAP is a low stakes national achievement tests developed and administered by CMEC across the 13 provinces and territories of Canada. It is administered every three years to 13 years old grade 8 junior high school students. The exams cover three main areas: reading, mathematics and sciences, with every round of examinations focusing on one area. In this paper, we analyze the data collected from two rounds of this exams, i.e. 2007 and 2010 PCAP[1]. See Table 1 for a brief summary.

Along with the exams, CMEC administered questionnaires on students, teachers, and schools. These incorporate the

[1]This Data is available upon request from:
`http://www.cmec.ca/240/Programs-and-Initiatives/Assessment/Pan-Canadian-Assessment-Program-(PCAP)/Overview/index.html`

Year	2007	2010
Major area	Reading	Mathematics
Minor areas	Math & Science	Science & Reading
Schools	15,00	1,600
Target	13 yrs old	Grade 8
Students	30,000	32,000
Major EN/FR	15,000/5,000	24,000/80,000
Minor EN/FR	7,500/2,500	24,000/80,000

Table 1: Summary of the PCAP dataset. For example, the 2007 PCAP focused on reading (which incorporated comprehension, interpretation, and responses). About 30,000 13-year-old students took the test, from more than 1,500 schools. For the major components, i.e. reading, 15,000 students took the English version, and 5,000 students took the French version. For the minor topics (mathematics and science components), 7,500 students took the test in English and 2,500 took it in French. For the 2010 PCAP, students answered questions in all three domains in the same language, with approximately 24,000 responding in English and 8,000 in French. This is in agreement with the general demographic of Canada, since one could say that roughly a third of Canadians are French speaking.

general background information, as well as reading habits and preferences, parental involvement, and types and frequency of reading.

The main component in the PCAP exams is the design of the assessment, questions of exams and questionnaires. Much effort is put into this design. Special care is put into develop questions that cover the intended subject domains, that includes different level of performance, considers curriculum of different jurisdiction, and at the same time ensure a fair and equal bilingual test development. Special care is also given when it comes to adequate and proper sampling of students and schools from across Canada. The analysis of the collected data however is limited to basic statistics, such as population mean per province, per gender, etc. For example, two of the main findings in the 2010 report by CMEC [2007] are that overall, for the performance of reading:

- "Female students had higher achievement than male students in both 2007 and 2010".
- "The difference between female and male students in 2010 was greater than it was in 2007".

This study tries to connect all the variables that would have an impact on the test results from a big picture perspective. We are in particular interested in the new insights obtained with the help of data mining techniques. The results presented in this paper are from our preliminary analyses and serve as motivation for future investigations. We discuss the current and future line work after the conclusion.

4. DATA MINING ON PCAP DATA

The PCAP data, as most real world data, is complex. It includes considerable amount of missing values, data inconsistency and heterogeneity. Here we first discuss how to tackle these challenges.

4.1 Data Cleaning

There are different data analysis toolkit which include built-in functions for dealing with complexity of data (missing values, inconsistency and heterogeneity). Here, we use a Python package called Pandas. Pandas[2] is a Python package designed specifically for real world data analysis. It supports tabular data and heterogeneous columns such as those in SQL or Excel. Which is a good fit here as the PCAP data is recorded in form of tabular data in SAS and Excel format. Here we first elaborate the missing values and inconsistent values in the PCAP data and how they are dealt with. The cleaned data is stored and used in further analyses.

4.1.1 Missing Values

Many real world data come with missing values. In PCAP also a large portion of data is missing, i.e. partially and/or not answered questions in the exams or questionnaires. The simple way to treat missing values is to exclude them or set them to a default value e.g. 0, which seems to be the practice used by the original statistics on the 2007 PCAP data. In the 2010 statistics, they used data imputation to predict the missing values based on a multiple regression analysis.

Pandas has special built-in functions to treat missing values. The "missing values propagate naturally through arithmetic operations between pandas objects". And each analysis handles them differently. For example they will be treated as zero when summing and excluded in correlation computation. They can further be interpolated using different approaches including 1-D, multivariate and spline interpolation.

For the PCAP data, the missing values are coded differently throughout different columns or even in one columns [3]. We simply clean and convert these to the Pandas Missing value. This way they are marked as missing and will be dealt with in the appropriate manner during further analysis.

4.1.2 Inconsistencies and Duplicates

Another challenge in real work application is the inconsistencies and redundancies in data. We clean the PCAP data, by carefully examining such cases and dropping redundant columns and/or those that have bad formatted values [4].

Moreover, there are missing values that should be treated as values and vice versa. Particularly, a performance grade of 0 is unlikely and it should be assumed to be missing, it is not a result of 0. Therefore we mark 0 performances as missing [5]. At last, the missing values of column program are filled with 'Not I' as oppose to 'I'. This makes sense since the program has only value 'I' for french immersion programs and value is missing for those that are not.

[2]http://pandas.pydata.org/pandas-docs/stable/

[3]The following values are treated as missing: '9,999', '999999', '99999', '9999', '999', '99', '9', '8' ,", '.' ,'not used' ,'Not used' , 'N/A' ,'Non-utilise', 'Non-assigne', 'Cahier non utili'.

[4]Specifically, the column 'TEACHER' seems to be a duplicate (of columns 'TEACHER_1' and 'TEACHER_2') and has unformatted bad values, and therefore is dropped. Also column 'frenchimm' is also dropped as it is a duplicate column, and has same information as the another column i.e. 'program'.

[5]That is values in columns: 'science','math', 'science500','math500', 'mathsciwt','Read', 'readwt', 'proficiency', 'READ500','comp500', 'interp500','resp500'

ID	Description	Values
S1_01	Gender	Male/Female
S1_02	Grade	Grade6..10
S1_03	Born in Canada	Yes/No
S1_04	Age they came to Canada	<5<10<
S1_05	Language used in home	EN/FR/CA/O
S1_06	Aboriginal ancestry	Yes/No
S1_07	Amount of books at home	scale 1..5
S1_08	Mother education level	scale 1..7
S1_09	Language used un school	EN/FR/CA/O
S1_10	French Immersion?	Yes/No

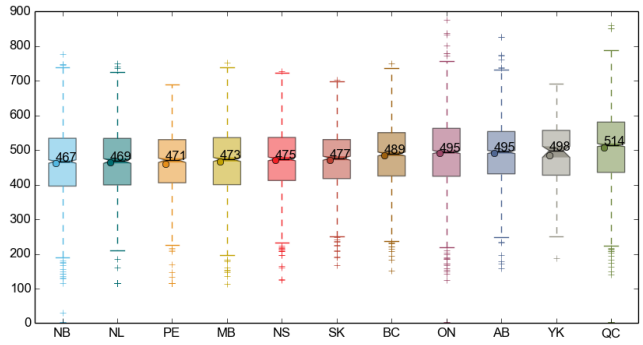
Table 2: Demographic Questions in Students Questionnaires. This table shows the 10 first questions of the questionnaires and their corresponding values.

4.2 PCAP 2007 Dataset Description

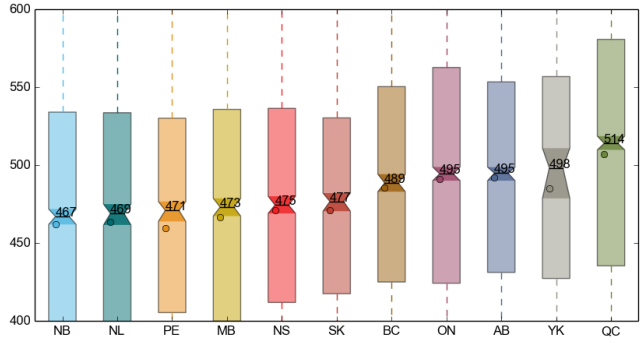
PCAP 2007 dataset includes 30022 students and 172 features which are their exams performance, their answers to the survey questions and information on their schools, and teachers. The school and teacher data further includes demographic information on them obtained from the questionnaires they answered such as total student enrolment, grade levels taught, language used in administration ('Anglophone' v.s. 'Francophone'). In Table 2, we highlight some of the features in the surveys, and their corresponding column number in the original data[6].

4.3 Descriptive Statistics

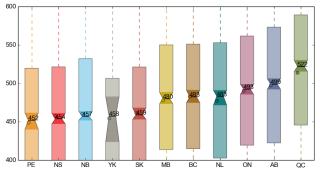
Having data cleaned and loaded, we can easily compute different statistics. Figure 1 for example compares performance of students across different provinces using box plots. Box plots are descriptive statistics used when there is no prior information about the distribution of the underlying population. We can see that Quebec students have overall a better performance in Reading and Mathematics, whereas Alberta students perform better than other provinces in science[7]. This complies with the findings previously published by the CMEC. However the overall ranking we have for provinces is different than the statistics of CMEC. For instance, in the results published by CMEC, Ontario is the second best province after Quebec in Reading and Mathematics, whereas we have it ranked fourth in Reading and third in Mathematics. The difference is possibly due to the fact that our ranking is based on the median instead of the average. Median is in general more appropriate than average if we don't know the underlying distribution of the data and also it is less affected with the outliers in the data. We can see from the box plots that there are indeed more outliers in the Ontario compared to other provinces[8]. Beside different rankings, the difference between provinces is in general less significant in our statistics. We have computed the confidence intervals using bootstrapping method, which are marked with notches in the box plots. For example we



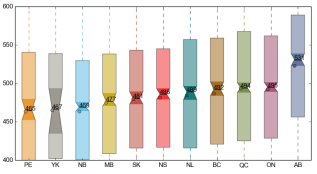
(a) Reading Performance



(b) Reading Performance Zoomed



(c) Mathematics Performance



(d) Science Performance

Figure 1: Boxplots for performance of students grouped by their jurisdictions. The red lines show the median of the scores in each province where the boxes denote the quartile of the population (25%-70% of data points are placed inside the box). Filled circles show the average, and the notches/darker areas represent the 95% confidence intervals.

find that there is no significant difference between Reading performance of students in Ontario and Alberta and Yukon, whereas Ontario is deemed significantly outperforming the other two according to CMEC [2007](pages 19,35-36).

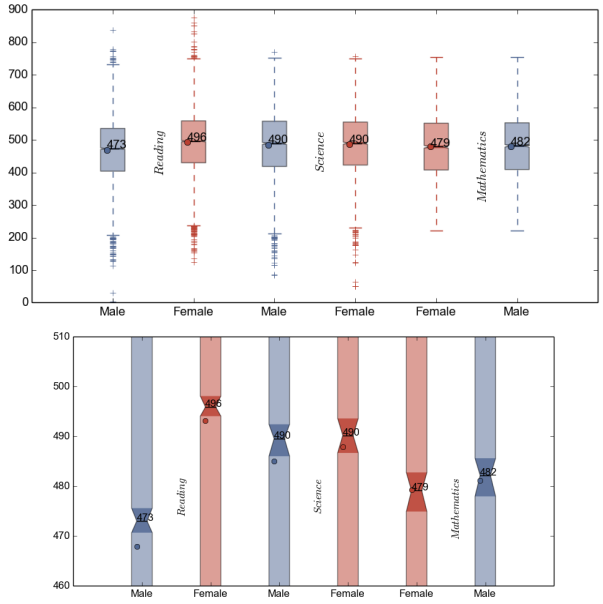
We compare the performance in similar manner for two other factors: gender and language. Figure 2a compares performance of female and male students. And Figure 2b compares performance of English and French speaking students, whom took the test in the corresponding language. In Figure 2a we can see that female students have slightly better performance on Reading. However the difference is not significant in Math and Science. This is similar to the findings published by CMEC [2007]. With this plots, we can also compare the overall shape of the two populations, not only their average performance. For example we can see that the low achieving outliers are more common in science than mathematics. And that high achieving exceptions in reading are more common between female students. In Figure 2b, we see that students who took the French version of the test performed better in Reading and Mathematics, how-

[6]The complete survey questions and reports are available online at `http://www.cmec.ca/docs/pcap/pcap2007/StudentQuestionnaire_en.pdf`

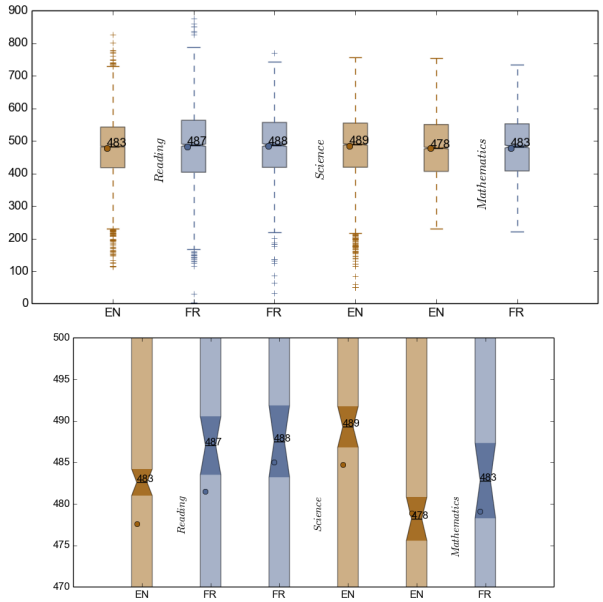
[7]We report the the weighted scores provided from CMEC by Item Response Theory. The trend is the same with raw scores. Here, we present the weighted scores for consistency with the original statistics published by CMEC.

[8]Outliers in the box plots are marked by a plus sign.

ever the difference is not significant. We could further refine and combine these comparison by different factors such as province and language (Figure 3b) and gender and language (Figure 3b).

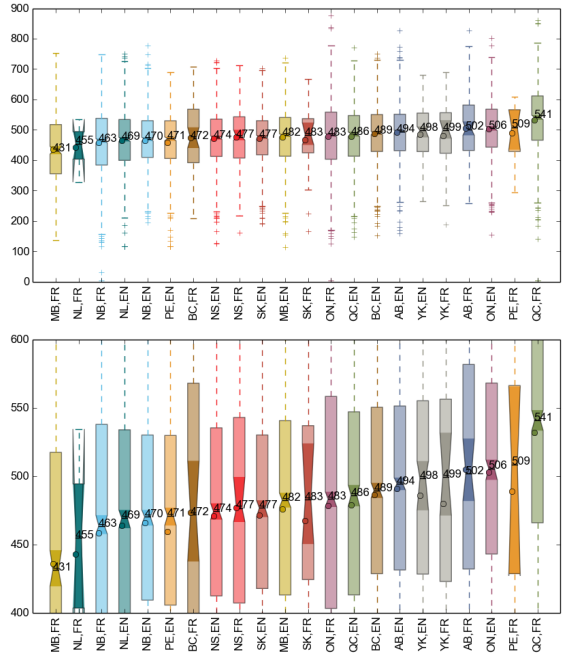


(a) Comparison of Performance between Male and Female, zoomed plot at bottom for significance comparison

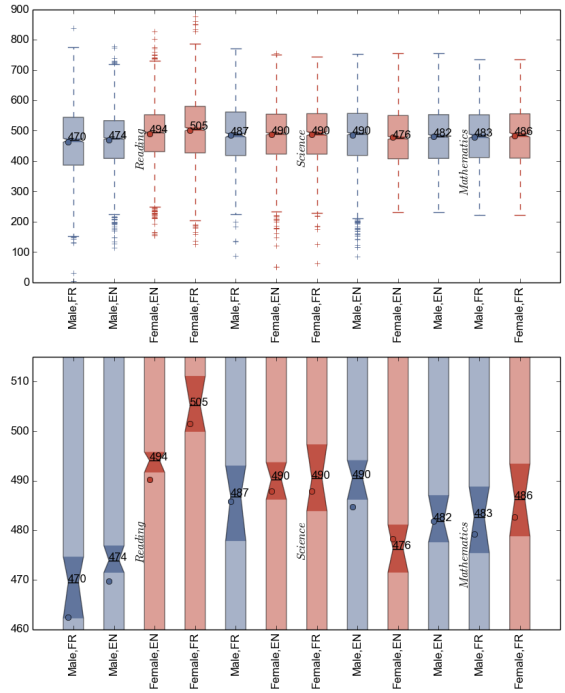


(b) Comparison of Performance between English and French, zoomed plot at bottom for significance comparison

Figure 2: Comparison for reading, science and mathematics between two languages and two genders. For the technical information about this plot please refer to the caption of Figure 1.



(a) French v.s. English per province



(b) French v.s. English per gender

Figure 3: Statistics for comparing reading performance of French and English narrowed by gender and province. From Figure 3b, for example, we find that the higher performance of Quebec students in reading is limited to French speaking students. On the other hand, in Figure 3b, we see that both English and French speaking Female students significantly perform better on Reading compared to their Male counterparts, whereas the gap is sharper among French speaking students.

4.4 Features Ranking and Clustering

We have studied and compared three factors, i.e. gender, language and jurisdiction. There are however many other factors that could be considered. In particular, the survey questions incorporate about 160 different demographic characteristics. For examples of such factors please refer to Table 2. The statistics regarding each of these factors could be studied individually. This is however time consuming, and a method is desired that would automatically detect the important factor(s).

For identifying important factors automatically, we compute all the pairwise correlations between different features. Figure 4 shows the obtained correlation matrix. Here we see some strong correlation between subsequent questions, i.e. small squares along the diagonal. This is expected since they correspond to the same aspect, for example S2_02A asks students if they enjoy reading, S2_02B ask if they read only when they have to, and S2_02C asks if they like to receive a book as a present, etc.

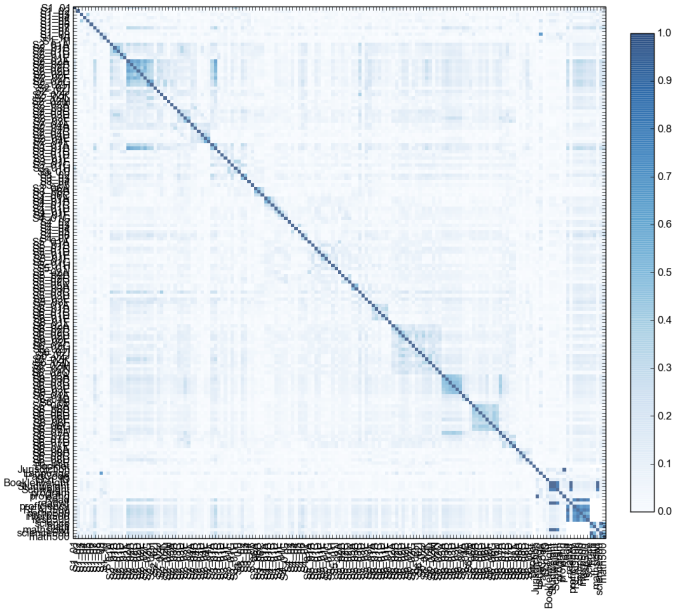


Figure 4: (Pearson) Correlation of different features. Missing values are excluded by Pandas when computing the correlation. However we filled the missing values for two sets of questions that include select those that apply. Since a missing value in this case means the item does not apply, rather than missing. Also we encode the categorical variables, before computing the correlations, using the factorize functionality in Pandas. The other technical point in generating this figure is that the absolute value of the correlations is considered here, so a strong correlation could be either positive or negative.

We then convert this correlation matrix to an undirected weighted graph, in which nodes are our features (S2_02A, S2_02B, ...) and each weighted link between two features corresponds to their correlation. Figure 5 visualizes the resulted correlation graph.

Using this graph representation, we rank and cluster the features. In more details, we first use the PageRank algorithm to rank the nodes in the graph, which are our features. The obtained ranking highlights the important features, which is reflected in the node sizes and also their position. In more details, the more important nodes are bigger and are also positioned closer to the center in the dual circular layout.

We further cluster the features based on their correlation, and group the highly correlated features together. The clustering is based on the connection between the features, which is performed using a modularity maximization network clustering algorithm described in Blondel et al. [2008]. In Figure 5, the node colours reflect the obtained groupings, i.e. nodes that belong to the same group have the same colour.

Using this ranking and clustering results we can infer the important features, in terms of the PageRank in the correlation graph, that are highly correlated with the students' scores. Here for example, we see that the students' answers to the following survey questions have a high influence on their performance on ...

- Reading (Blue Cluster):

S2_02A Whether and to which degree they enjoy reading.

S3_01B How much time they spend on reading for enjoyment and/or general interest outside of the school hours.

S2_02D Whether and to which degree they think reading is a waste of time.

S2_02F Whether and to which degree they enjoy going to a bookstore or library.

S3_01A How much time they spend on outside-of-class reading for their courses.

- Math and Science (Red Cluster):

S4_05 Whether they are given a rubric when they start an assignment in the English Language Arts classes.

S4_03 If they know what a scoring rubric is for marking tests or assignments.

S5_03B How much they think the reading they do in school for English Language Arts classes is more appropriate for boys than girls.

S5_03A How much they think the reading they do in school for English Language Arts classes is more appropriate for girls than boys.

S3_01C How much time they spend doing sports or other school and community activities outside the school hours.

The latter factors are quite interesting. We see that the survey questions regarding art and sport activities have a strong correlation with students performance on Math and Science modules. Here, we considered the absolute values of the correlations, so we can not infer whether the the relation is positive or negative. In fact such conclusion needs a much more thorough analysis, which is part of the future work for this study. In particular, we need to make the connections not based on the correlations of features, but based on the a prediction accuracy. We also need to built and incorporate an ontology for the questions and answers in the surveys, in order to infer the student answers, and detect positive and negative factors.

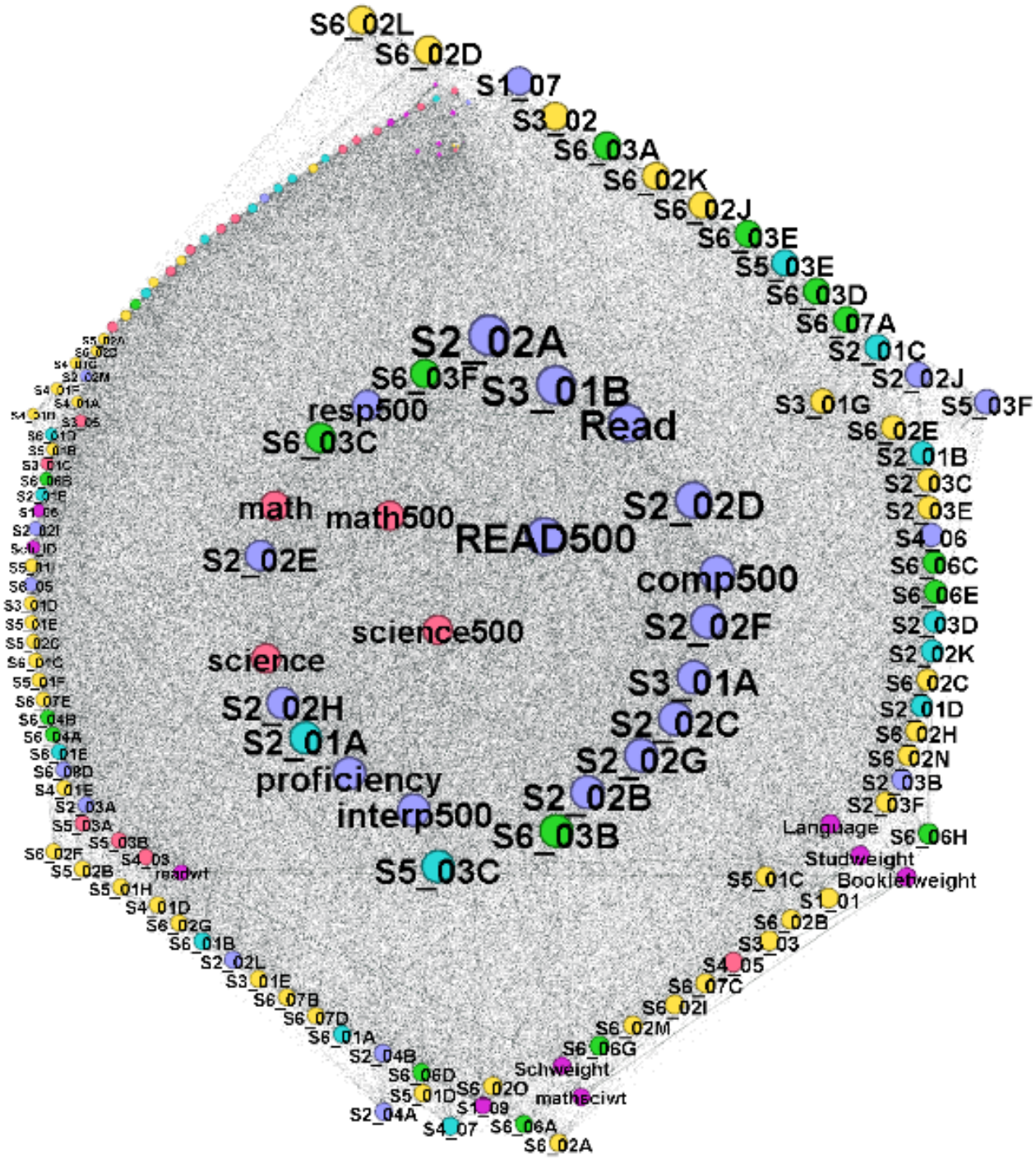


Figure 5: Graph representation of the correlation matrix. Here nodes are our features, (i.e. survey question, demographic information on students and also their performance scores). Size of the node corresponds to its weighted PageRank in this correlation graph. This visualization is generated using Gephi toolbox. The layout used is a Dual Circle Layout, where the ranking of the nodes is used in positioning them in the circles, and the 25 highest ranked features are placed in the inside of the circle. Nodes are also colour coded based on the cluster/group they belong to. Grouping is obtained using a weighted version of the Modularity algorithm for clustering graphs. Here we see five main groups of correlated features. The blue cluster includes reading scores and the features correlated with it, whereas the red cluster includes the reading and math scores and their correlated features. The other factors grouped separately are showing different highly correlated features. For example the yellow cluster is related to the perseverance and hard-work. In particular, S6_02L asks if they re-read the difficult parts in order to understand the text, S6_02D asks if they try to make connections between what they read and what they already know, S3_02 asks how much time they spend on doing homework, and S3_01G asks if they use computers for school works.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a case study on the analysis of large scale assessment test data. The data we studied is from low-stake nationwide exams in Canada; and as such it provides a good measure of students' true achievement at the grade 8th level in reading, math and sciences. By taking off the stress caused by high stakes testing, the PCAP focuses on the "genuine" abilities of the students. The main contribution of this paper is showcasing possible analyses that could be performed on this type of data; thus providing stakeholders with insightful information from overlooked connections. In our case of the PCAP data, for example, the amount of effort for assessment and test organization design is not comparable by the effort put into choosing and justifying the type of analyses performed.

We would like to mention that our study is a work in progress and there are many further analyses that we have planned to perform as our future works. These include: changing the correlation dependencies to the prediction accuracy, using different feature selection methods to find the features' importance, etc. Moreover, we are also aiming to perform an in depth item level analysis of the exam questions to detect irregular questions, and to rank questions based on their difficulties. The other further work for our study is incorporating other sources of information, i.e. the demographic information regarding teachers and schools, and also comparing the results with the PCAP exams from other years.

Lastly, we are planning to add the results of the PISA[9] tests for Canadian students and comparing it to the PCAP. CMEC administers the PCAP to 13 years old as a national achievement indicator and to these same students, it administers the PISA two years later when they are 15th years old. The focus of both assessment programs is the same: for instance the PCAP 2010 focused on mathematics, and the PISA 2012 was also focused on Mathematics. With such growing roles of nationally and internationally used battery of assessment used for quality assurance of global education, we foresee the potential of data mining with a longitudinal data tracking the students' performance in such situation.

References

RSJD Baker. Data mining for education. In B. McGaw, P. Peterson, and E. Baker, editors, *International encyclopedia of education*, pages 112–118. Elsevier, Oxford, UK, 3 edition, 2010.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

CMEC. Report on the assessment of 13-year-olds in reading, mathematics, and science. Technical report, Council of Ministers of Education, Canada, 2007. URL `http://www.cmec.ca/Publications/Lists/Publications/Attachments/124/PCAP2007-Report.en.pdf`. PCAP-13 2007.

[9]PISA is the Programme for International Student Assessment, developed by the Organization for Economic Development as an international measure for students scholastic performance in mathematics, reading and sciences. Over 60 nations take part in the test.

John Dewey. Science as subject-matter and as method. *Science*, 31(787):pp. 121–127, 1910. ISSN 00368075. URL `http://www.jstor.org/stable/1634781`.

Veronica Diaz and Malcolm Brown. Learning analytics: A report on the eli focus session, 2012.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

Janice D Gobert, Michael Sao Pedro, Juelaila Raziuddin, and Ryan S Baker. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4):521–563, 2013.

Ronald K Hambleton and R. W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3): 38–47, 1993. URL `http://ncme.org/linkservid/66968080-1320-5CAE-6E4E546A2E4FA9E1/showMeta/0/`.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of item response theory*. Sage, 1991.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

Richard A Huebner. A survey of educational data-mining research. *Research in Higher Education Journal*, 19, 2013.

R. L. Linn and N. E. Gronlund. *Measurement and Assessment in Teaching*. Upper Saddle River, NJ: Prentice Hall, 8 edition, 2000.

Frederic M Lord, Melvin R Novick, and Allan Birnbaum. Statistical theories of mental test scores. 1968.

Sharan B Merriam, Rosemary S Caffarella, and Lisa M Baumgartner. *Learning in adulthood: A comprehensive guide*. John Wiley & Sons, 2012.

James W Pellegrino. *The evolution of educational assessment: Considering the past and imagining the future*. Educational Testing Service, Policy Evaluation and Research Center, Policy Information Center, 2004.

W James Popham. Instructional sensitivity: Educational accountability's dire deficit. *Phi Delta Kappan*, 89(2): 149–155, 2007.

W James Popham. *Transformative assessment*. ASCD, 2008.

Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

George Siemens and Ryan SJ Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.