

Application of Learning Analytics to Investigate Factors in Determining Reading Proficiency¹²

Dr. Samira ElAtia
The University of Alberta
Edmonton, Alberta, Canada

Dr. Ahmed Hammad
Sharjah University
Sharjah, United Arab Emirates

Abstract

*Author Malcolm Gladwell argues in both his bestsellers books *The Tipping Point* and *Outliers* that extreme intelligence is not exclusively responsible for success. Neither is giftedness. One could be extremely gifted, but if the surrounding environment is not supportive, no greatness would be achieved and no endeavor would be successful. Most importantly, Gladwell singled out perseverance as a key element of success and as being far more important than the “intelligent quotient” alone. Can the same thing be applied to school subjects? More specifically, is reading linked to personal choices and preferences? Or can the environment have an effect on reading competence? If so, what factors can sharpen the “gift” of reading?*

For the last 5 decades or more, these questions have been at the heart of much research around the world. In this study, we used Learning Analytics to study data in novel ways within a big data perspective. Learning Analytics (LA) is defined as the “use of data, statistical analysis, and explanatory and predictive models to gain insights and act on complex issues about the learners “by Diaz & Brown (2012). We used LA to explore factors that influence performance in reading among school children. Using a large data set (approximately 9000 subjects) from across Canada, we aimed to find out what elements can predict excellence in advanced reading. The data is from the PanCanadian Assessment Program (PCAP), a national achievement test administered by the Council of Ministers of Education, Canada. LA allowed us to go beyond traditional statistical analysis; we were able to use data mining techniques to advance our understanding of reading competence from a big picture perspective. Instead of using only the individual results, we wanted to find what other factors influence reading beyond personal strength or intelligence alone, or more appropriately, we want to find out what contribute to building personal strength in reading.

Along with the test results, the PCAP data include responses to three extensive questionnaires: from students/parents, teachers, and school administration. The extensive data, both the surveys and the test results, gave us a unique opportunity to delve into reading competence. Reading is directly linked to Literacy, identified by one of the core competencies for elementary education in Canada. Our findings were interesting since we were able to single out new factors that were not significant in other studies that relied on classical statistics methods. As an educational research tool, LA comes with new innovation at how we look at Data from elementary schools

¹ This study has been funded by a grant from the Social Sciences and Research Council of Canada, and a grant from the Killam Research Funds of the University of Alberta.

² This paper is part of a large project and a continuation of several studies conducted over a period of 4 years (ElAtia et al 2012; ElAtia & Hammad 2011, and Fung & ElAtia 2015)

Reading at the intersection of Cognition, Literacy and Language Proficiency

Reading is no easy task. It is a complex task during which two elements are happening: a process of reading and a product of reading; from deciphering the written codes and the meanings of words and sentences to understanding messages and connotations. Both elements are intertwined and involve multiple tasks and subtasks. Regardless of whether it is in a first language or second/foreign language, reading is a linguistic, cognitive, and metacognitive task that requires particular skills for it to happen and according to Bachman (2000) is “the most extensively researched and most enigmatic of the so-called language skills.”

Reading is at the heart of Literacy. While generally for any native speaking students, reading is challenging, for the non-native language speakers the challenge is even more ‘challenging’ and it is important to understand the complexities of this group. The PCAP reading assessment protocol provides a unique holistic view of students’ performance since (a) it provides not only the results of the tests, but also the surveys from schools, teachers, and parents (b) it includes two language groups, French and English, with native and non-native speakers. Data for the PCAP provides us with a unique opportunity to understanding the intricate skill that is reading.

For Alderson (2000) there are two levels of understanding: one is the literal understanding of the text presented to the reader, and the other is the “understanding of meaning that are not directly stated in text... understanding of the main implications of text” (p.7). Gray (1960) made three distinctions:

- (1) linked to the literal meaning (reading the lines),
- (2) inferring meaning from text (between the lines) and
- (3) higher critical thinking meaning (beyond the lines).

In light of these distinctions alone, the results of the PCAP would be directly influenced by the students’ linguistic and educational background. On the one hand, we need to question if they have matured cognitively enough to be in the same reading age as the other examinees. Do they possess what Grabe (1991) refers to as the speed and automaticity of recalling words and linked words to meaning? Grabe lists the following six elements necessary for a fluent reader :

- (a) automatic recognition skills,
- (b) vocabulary and structural knowledge,
- (c) formal discourse structure knowledge
- (d) content/world background knowledge,
- (e) synthesis and evaluation skills and strategies, and
- (f) metacognitive knowledge and skills monitoring.

From a metacognitive perspective, previous knowledge of the text or familiarity with the topic of the reading texts (Fisher, 2012) may play a major role in the test performance. The fact that within the two language groups we identified as non-native language students (nonnative language learners) may have either linguistics difficulties or educational gaps, would play a major role in the development of the metacognitive skills necessary to understating texts. Abedi (2008) strongly advocates paying close attention to language proficiency when assessing for content among school students. In addition, items in the PCAP are culturally ‘Canadian.’ That is,

items in the reading passages were about a game of hockey among school children for instance, or about birds off the Bay of Fundi. Studies in differential item functioning did find out that previous knowledge of subjects in reading tests does influence test outcomes/performance (Recht, 1988). We wonder to what extent such factors have an impact on the students' performance in the PCAP.

The purpose of this study was to find out what we can learn from the extensive PCAP data if we used Learning Analytics and datamining to investigate links among the data from a big data perspective. We sought to find new connection, new clustering, and new outliers that are not necessarily detected by classical statistical targeted analyses. For our research question to be answered, we were privileged to have a wealth of Data in the PCAP.

Description of the Data

This study was conducted in the Canadian context where we used a large scale national assessment data. The Pan-Canadian Assessment Program (PCAP) is a Canada-wide achievement test that assesses performance of 13-year-old (Grade 8) students on their reading, mathematics and science skills. The PCAP is administered every three years, and each administration focuses on one major core subject and two minor subjects, alternating between mathematics, reading, and science: for instance, in 2007, the focus was on reading, in 2010, the focus was on mathematics, and in 2013 it was on sciences, most recently in 2016, the focus was on reading again. The administration of the PCAP overlaps with the 'Programme for International Student Assessment' (PISA); and it allows the Canadian government not only to conduct quality assurance across the country, it allows for international comparative studies, and extensive quantitative and qualitative investigations.

The PCAP is developed by the Council of Ministers of Education, Canada (CMEC) in an attempt to provide a large-scale assessment and reporting on students' achievement across Canada (CMEC, n.d.). The previous form of PCAP was the School Achievement Indicator Program (SAIP), which had been administered nine times between 1993 and 2004. In 2003, a decision was made by CMEC to replace the SAIP with the PCAP due to changes in national curriculums and the need to reflect on students' learning across Canada. The test was administered in both English and French to schools from across Canada.

There are two main purposes of the PCAP:

“(a) to help ministries and departments of education make informed decisions about education, and

(b) to determine how prepared 13-year-old students are for the academic challenges of high school (CMEC, n.d., p.1).”

Since the entire purpose of the PCAP is to ensure equality of student learning across Canada, in regards to preparation for high school entry, the CMEC (n.d.) stressed that results of the test should not be used to reflect students' individual performance, nor should results be used to compare schools and jurisdictions. To avoid the misuse of the test results, the PCAP results are not available to school boards, ministries of educations, schools, or teachers. Along with taking the assessment, students (and their parents) are asked to respond to a questionnaire targeting detailed information such as reading habits and attitude. Additionally, a teacher questionnaire, and a school questionnaire are also administered. The PCAP data thus becomes a very rich and substantial one.

To provide an idea of the scope of the data, the English part of the Reading assessment, 1281 schools and a total of 15,000, 13-year-old students participated. The results and the answers

to the questionnaire provide a unique opportunity to study students' achievement by looking at various elements that could contribute to its progress. For this study, we used data entries for the 7,500 students took the test in English and 2,500 took it in French, we focused on these entries because they are the most complete ones.

Educational data mining and Learning Analytics

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, Fayyadet 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 generated within educational contexts.

Educational Data Mining (EDM) applies data mining techniques to explore data originating from various 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 behavior 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.

For this study, we are transforming learning from reading data by looking at data from a different angle. We are studying 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.

Discussion of results

By being a low stakes exam, results of the PCAP are a window into the true reading abilities of these grade 8 students. By removing the stress of high stakes achievement tests, students performance is very close to their own reading abilities, it is more natural, more real. The first set of analysis we carried is ranking and clustering results, where we could infer the important features that are highly correlated with the students' scores. Below, the following survey questions have a high influence on the students obtained test results (i.e. performance):

- ***Reading clustering:***

- 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.

For reading cluster, the items above are found to be high predictors of students' performance in reading. Among these, the mothers' level of education is the highest factor in predicting students reading abilities.

Math and Science clustering

- 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 items above (see appendix for graphical displays) from the survey had strong correlations with students' performance in Math and Sciences: participating in arts and sports activities correlates highly with higher test results.

The use of a Canadian performance measure (i.e., the PCAP) allowed results to be applied in a multicultural framework. It could further help various stakeholders in assisting students to improve their reading abilities, especially for language learners and for bilingual contexts. In addition, parents and teachers could benefit from these results and improve students' reading environment and habit. For example, teachers may create more targeted opportunities for students to practice their reading skills, hence allowing them more exposure within the school for gaining reading experiences. Parents may also implement strategies at home for encouraging reading.

This was a first study of its kind to LA and DM on such data as the PCAP. Future research could investigate important factors for consideration in teaching reading. We plan to replicate the study using the 2016 PCAP data once it is made available to researchers. We also want to compare our results to confirmatory factor analysis studies that we would like to carry based on the initial reports that identified the times in the clustering analysis.

References

- Abedi 2008 Abedi, J. (2008). "Measuring students' level of English proficiency: Educational significance and assessment requirements." *Educational Assessment*, 13 (2-3).
- Abedi, J. (2004) *The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues*. *Educational Research*, 33(1).
- Alderson (200) Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Bachman (2000) Bachman, L. F. (2000). Preface. In J. C. Alderson, *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Baker, R. (2010) Data mining for education. In B. McGaw, P. Peterson, and E. Baker, editors, *International encyclopedia of education*. Elsevier, Oxford.
- CMEC (n.d.). *Pan-Canadian assessment program (PCAP)*. Retrieved from [http://www.cmec.ca/240/Programs-and-Initiatives/Assessment/Pan-Canadian-Assessment-Program-\(PCAP\)/Overview/index.html](http://www.cmec.ca/240/Programs-and-Initiatives/Assessment/Pan-Canadian-Assessment-Program-(PCAP)/Overview/index.html)
- CMEC. (n.d.) 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.
- Dewey, John. (1910) Science as subject-matter and as method. *Science*, 31(787).
- Diaz, V & M. Brown (2012). *Learning analytics: A report on the eli focus session*,.
- ElAtia, S., Ipperciel, D. & Hammad, A. (2012), "Implications and Challenges to Using Data Mining in Educational Research in the Canadian Context". *Canadian Journal of Education*, Volume 35, Issue 2: (June 2012)
- ElAtia, S. & Hammad, A. (2011), "Advancing Educational Research Through a Knowledge Discovery in Data (KDD) Model". American Educational Research Association Conference, in New Orleans, (May 2011).
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996) From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Fisher, D., Frey, N., & Lapp, D. (2012). Building and activating students' background knowledge: It's what they already know that counts. *Middle School Journal*, 22-31
- Grabe, W. (1991). Current developments in second-language reading research. *TESOL Quarterly*, 25(3).

- Gobert, D., Sao Pedro, M., Raziuddin, J., & Baker, R. (2013). From log les to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4).
- Gray, W. S. (1960). The major aspects of reading. In H. Robinson (Ed.), *Sequential development of reading abilities* (pp. 8–24). Chicago, IL: Chicago University Press.
- Han, J., Kamber, M. & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann
- Huebner, R. (2013). A survey of educational data-mining research. *Research in Higher Education Journal*, 19.
- Organisation for Economic Cooperation and Development. (2013, June). PISA in focus. Retrieved from [http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa in focus n29 \(eng\)--Final.pdf](http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa%20in%20focus%20n29%20(eng)--Final.pdf)OECD (June 2013).PISA in focus. Retrieved from [http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa in focus n29 \(eng\)--Final.pdf](http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa%20in%20focus%20n29%20(eng)--Final.pdf)
- Rabbany, R., Zaiane, O. R., & ElAtia, S. (2014, August). *Mining large scale data from National Educational Achievement Test*. Presentation at the KDD Workshop on Data Mining for Educational Assessment and Feedback, in conjunction with ACM SIGKDD 2014, New York City, NY.
- Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of test. *Journal of Educational Psychology*, 80(1).
- Romero, C. & Ventura, S. (2010). 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).
- Siemens, G. & Baker, R. (2012). Learning analytics and educational data mining: towards communication and col-laboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*.

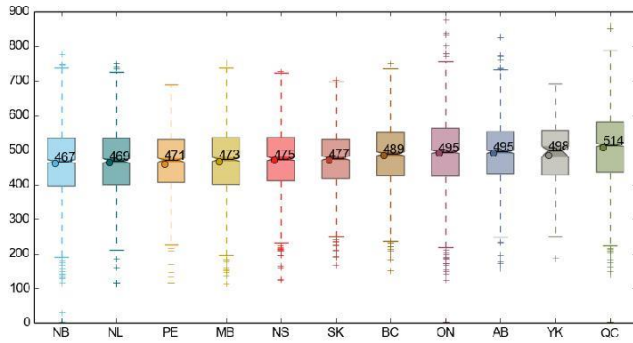
Appendices (The DM analysis that were carried out)

Demographic Questions in Students's Questionnaires

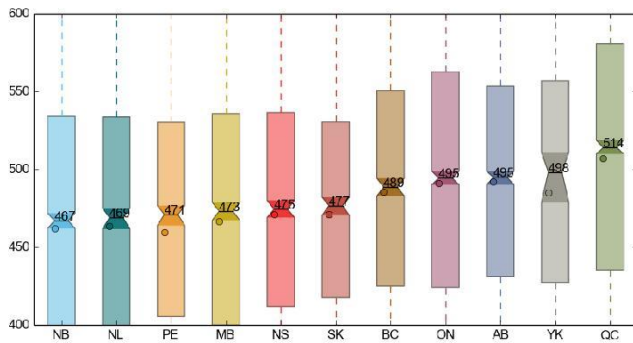
ID	Description
S1 - 01	Gender
S1 - 02	Grade
S1 - 03	Born in Canada
S1 - 04	Age they came to Canada
S1 - 05	Language used in home
S1 - 06	Aboriginal ancestry
S1 - 07	Amount of books at home
S1 - 08	Mother education level
S1 - 09	Language used un school
S1 - 10	French Immersion?

(The complete survey questions and reports are available online, and upon request from CMEC at http://www.cmec.ca/docs/pcap/pcap2007/ StudentQuestionnaire_en.pdf)

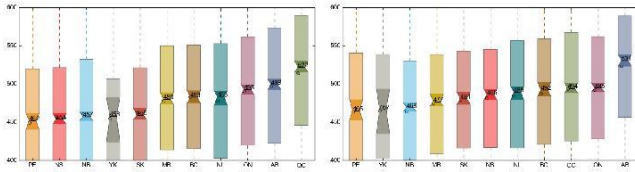
Boxplots for performance of students grouped by their respective provinces



(a) Reading Performance



(b) Reading Performance Zoomed

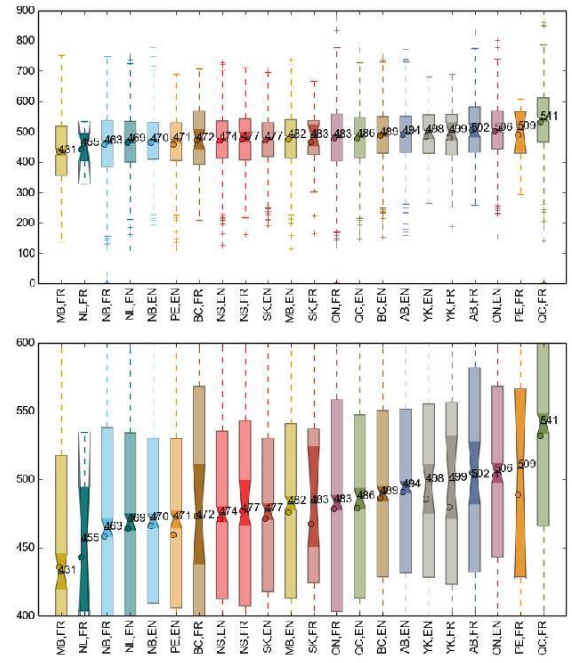
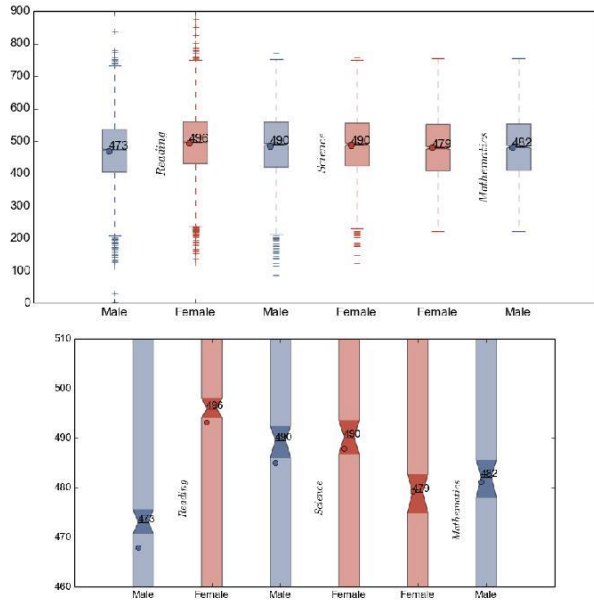


(c) Mathematics Performance

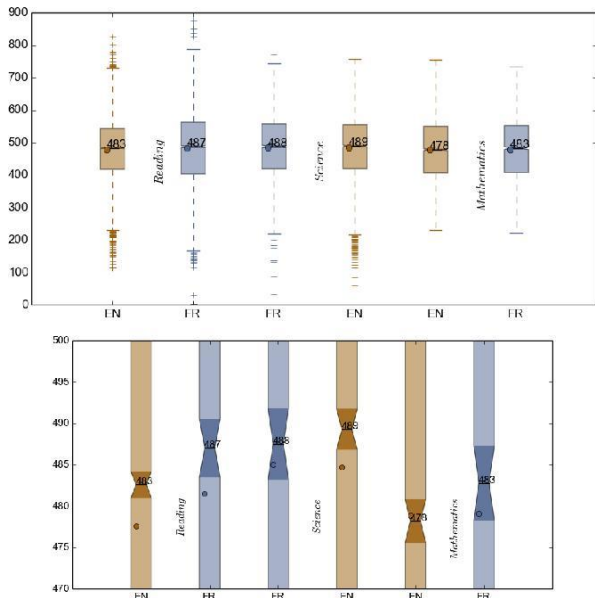
(d) Science Performance

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.

Comparison for reading, science and mathematics between two languages and two genders

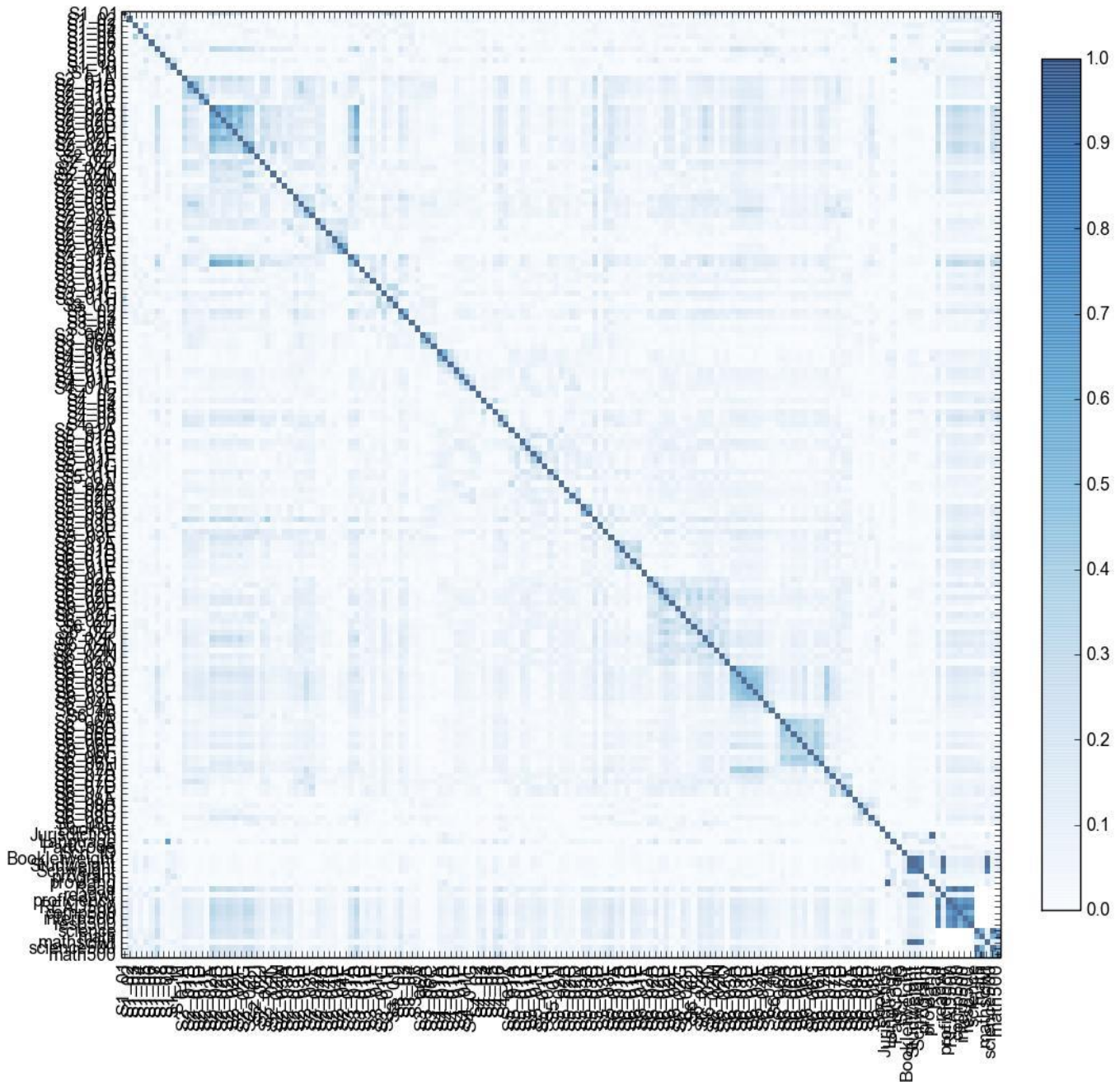


(a) Comparison of Performance between Male and Female, zoomed plot at bottom for significance comparison



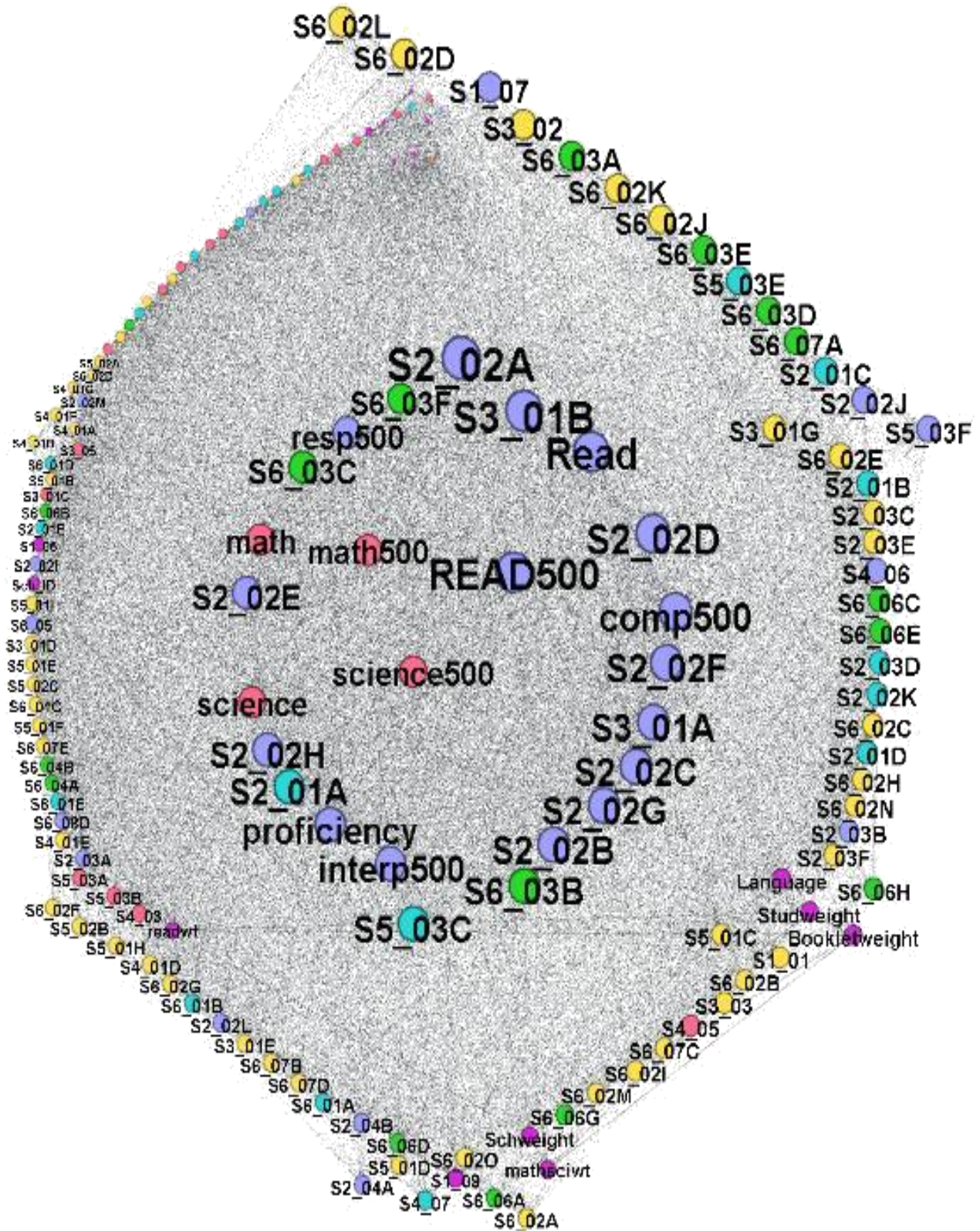
(b) Comparison of Performance between English and French, zoomed plot at bottom for significance comparison

Correlation of different features



Missing values are excluded by Pandas when computing the correlation. However we considered 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.

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 color 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 the 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.

بسم الله الرحمن الرحيم

تطبيق تحليلات التعلم لبحث العوامل المؤثرة على إتقان القراءة

يقول الكاتب مالكولم جلادويل في كتابيه الأكثر مبيعا "نقطة التحول" و "المختلون" أن الذكاء الشديد ليس وحده السبب الرئيسي للنجاح. وأيضا ليست الموهبة هي سر النجاح. يمكن للمرء أن يكون شديد الذكاء و موهوبا و لكن في حال عدم وجود بيئة مساعدة و داعمة فلن يحرز النجاح المنشود. حدد جلادويل بكل وضوح أن المثابرة هي العنصر الأساسي في إحراز النجاح و المثابرة أهم كثيرا من الذكاء أو الموهبة. هل بإمكاننا تطبيق نفس الفرضية علي قدرة الطلاب علي استيعاب المقررات الدراسية؟ و بتحديد أكثر: هل إتقان القراءة هو اختيار شخصي للطلاب أم أن البيئة المحيطة تلعب دورا مؤثرا في ذلك؟ و بافتراض صحة هذه الفرضية، فما هي هذه العوامل التي بإمكانها تحسين ملكة القراءة لدى الطلاب؟

خلال العقود الخمسة الأخيرة، تم طرح هذه الأسئلة في كثير من الأبحاث العلمية علي مستوى العالم بأسره. في هذه الدراسة قمنا باستعمال تحليلات التعلم من كمية كبيره من البيانات بطريقة جديدة. طبقا لدياز و براون (2012) يمكن تعريف تحليلات التعلم بأنها "استخدام البيانات و التحليلات الإحصائية و النماذج التنبؤية و التفسيرية لاستنباط المعرفة الكامنة في البيانات و التعامل مع الأمور المعقدة الخاصة بالطلاب".

قمنا باستعمال تحليلات التعلم للبحث عن العوامل المؤثرة علي مستوى إتقان القراءة بين أطفال المدارس. قمنا أيضا باستخدام كمية كبيره من البيانات (حوالي 9000 سجل بياني) من كل أنحاء كندا لمعرفة العناصر التي يمكنها التنبؤ بتميز الطلاب في القراءة المتقدمة. تم الحصول عل سجلات البيانات من "برنامج التقييم لعموم كندا" و هو امتحان وطني تتم إدارته عن طريق مجلس وزراء التعليم بكندا.

تحليلات التعلم مكنتنا من الحصول على نتائج أفضل من التحليل الإحصائي التقليدي و باستعمال تقنيات التنقيب عن المعرفة في سجلات البيانات تمكنا من الوصول لفهم أفضل للعوامل المؤثرة علي إتقان القراءة من ناحية الصورة الكلية.

وبدلا عن التركيز على النتائج الفردية، قمنا بالبحث عن العوامل المؤثرة علي إتقان القراءة خارج نطاق الذكاء أو الموهبة الشخصية أو بالأحرى العوامل التي تؤدي إلى تقوية مهارات القراءة المتقدمة.

و بالإضافة الي نتائج الامتحانات، احتوت سجلات البيانات على ردود الطلاب / ابائهم و المعلمون و مديرو المدارس على ثلاثة استبيانات شاملة. أتاحت لنا البيانات واسعة النطاق التي حصلنا عليها من كل من الامتحانات و الاستبيانات فرصة فريدة للخوض في دراسة العوامل المؤثرة علي إتقان القراءة.

معرفة القراءة و الكتابة واحدة من أهم الكفاءات الأساسية اللازمة للتعليم الابتدائي في كندا. ساعدت نتائج بحثنا على الوصول لعوامل جديدة مؤثرة علي إتقان القراءة لم تكن ذات أهمية في الدراسات التي أجريت من قبل باستخدام الطرق الإحصائية الكلاسيكية. إن تحليلات التعلم تقدم لنا وسيلة بحث تربوية جديدة للاستفادة من سجلات البيانات المتوفرة في المدارس الابتدائية.

د. سميرة العاطية

د. أحمد حماد

