Identifying key visual-cognitive processes in students' interpretation of graph representations using eye-tracking data and math/machine learning based data analysis

Enrique Garcia Moreno-Esteva [1], Sonia White [2], Joanne Wood [2] and Alexander Black [2]

[1]U. of Helsinki, Dept. of Teacher Edu., Helsinki, Finland; enrique.garciamoreno-esteva@helsinki.fi

[2]Queensland University of Technology, Brisbane, Australia; sl.white@qut.edu.au

We present a mathematical and computational analysis, partially based on machine learning techniques, of the optical tracks obtained during a graph interpretation task which allows us to identify when the problem solver succeeds in solving the problem with a fair degree of accuracy, and helps to understand the visual-cognitive processes at work during the problem solving task.

Keywords: Graph interpretation, Eye-tracking, Machine learning, Mathematics education, Gaze metrics

As a way of introduction: about the task and machine learning.

Eye-tracking is quickly becoming an established technique for investigating cognitive processes involved in the learning of mathematics and other subjects (Lai, et al., 2013). Unfortunately, the analysis of eye-tracking data is difficult and laborious, often involving frame by frame analysis (Garcia Moreno-Esteva, Hannula & Toivanen, 2016). We partially overcome this difficulty here, with the use of machine learning and other mathematical techniques. Using a desktop eye tracking system, children completed a mathematics problem that incorporated a bar graph. The optical tracks and the accuracy of the response are analyzed in order to understand how a child "reads a graph". We are trying to gather from our data and its analysis, a story of what happens when several children are confronted with such a task. What do they look at? Do the gaze patterns influence the success or accuracy when responding to the task? With this information we may be able to more reliably infer the cognitive processes completed by children.

The problem-solving task.

In Brisbane, Australia, a group of 113 children (mean age 8.67 years), all in the second half of year 3 in school, completed the graph problem solving task. As part of a larger project, children completed a series of eye tracking tasks (reading, mathematics) in a quiet room near their classroom. The mathematics tasks included odd-even judgement, magnitude comparison, and problem solving tasks: interpreting a bar graph and navigating a coordinate grid. The focus of this presentation is the graph problem solving task. This task was designed based on the Grade 3 Australian Curriculum Mathematics where Grade 3 children are expected to be interpreting and comparing data displays (ACARA, 2016). A similar graph interpretation task features in a Grade 3 Australian standardized achievement test. The children were shown the following: a) a bar-

chart, where the height of each bar indicated a the number of hours worked by Sarah during a given week; b) a labeled coordinate system, where the x-axis had the week number labels, and the y-axis had numbers corresponding to hours; c) a sentence indicating Sarah's hourly wage; d) another sentence indicating the task to completed related to Sarah's wages in Week 3. Curcio (2010) describes a sequential framework for children's data comprehension, this framework includes; *understanding, interpretation and prediction* with data. The current graph task required each child to *understand* and *interpret*: reading the question and basic details of the graph (*understanding*), and then reading between the different elements of information (*interpretation*) in order to complete the computation and arrive at the correct solution for Sarah's Week 3 earnings. A Tobii eye-tracker operating at 300 Hz recorded the locus of focus of their eyes throughout the activity - including the initial *understanding*, and steps involved in *interpretation*. The threshold for fixations was set at 100 ms (Tobii Technology, 2014). It was hoped that the eye tracking information (fixations and saccades) might shed light on the different cognitive steps involved in the task. Initial qualitative evaluations of the eye movements demonstrated children who did not progress past the first *understanding* stage, as they did not identify the question being asked or relevant information on the graph. Other children were able to *understand* the task and progressed to specific *interpretation* of relevant information - with a variety of behaviors demonstrated. For example, some children had high numbers of fixations and saccades around relevant areas, whereas others had fewer and longer fixations on relevant areas. These initial qualitative observations were systematically investigated using machine learning techniques.

The data included 113 optical tracks (for purposes of the forthcoming discussion, the *inputs*), and 113 answers (the *outputs*), considered as correct (1) or incorrect (0). The optical tracks consisted of sequence of pairs, each pair included the duration of a fixation in milliseconds (ms), and the location of the fixation. The optical track information can be visualized as a video (or a static picture) in which the fixations appear as a sequence or red dots that have a size proportional to the duration of the fixation, and which are connected by lines to neighboring fixations.

After inspecting the optical track videos, it was evident that it would be difficult to disentangle patterns of visual processing that might reveal cognitive processing of different children. It was decided that further mathematical/computational analysis of the data might provide further insight. Since the nature of the input data is sequential, classifying the optical tracks and test results (inputs and outputs) with a Markov model based machine learning technique was selected as an appropriate analytic method.

A word about machine learning techniques.

The proprietary algorithm (*Mathematica*'s Classify function) was used to do the machine learning analyses, using a Markov model method (Wolfram Language and System Documentation Center, 2016). In this analysis you select a subset of the sample (input – optical track - and output – result - data) to analyze (*classify*) with the machine learning algorithm. From that analysis a *classifier* is then used on all the inputs (optical tracks) to predict the outputs (0 or

1, incorrect or correct). The predicted outputs from the classifier are then compared to the real outputs, and the percentage of correctly classified outputs can be calculated (some examples are provided in subsequent sections).

Our research question.

Our research question is simply, what can we learn or infer about cognitive processes related to the graph interpretation task described with mathematical/computational/machine learning based analysis techniques of the eye-tracking data, and maybe, could these techniques be of further help in analyzing the data pertaining to other well defined mathematics problem solving tasks?

Our techniques are general, in that they can easily be applied to other eye-tracking data consisting of a sequence of fixations given by the coordinates and the durations of the fixations as inputs, and a set of two or even more categories as outputs. We hope to make the programs available to other researchers wanting to undertake this kind of analyses at a later stage or our research.

The analyses and corresponding results.

In this section we will describe three kinds of analysis for which we obtained encouraging results. Other possible analyses will be discussed in a later section pertaining to directions of future work.

We partitioned the visual stimulus (the graph on the screen; Figure 1) into areas of interest (AOIs), where the most critical areas of interest are labeled as A1 (wage information), A2 (week number), A3 (week 3 bar), and A4 (number region containing the number of hours corresponding to week 3), and other areas of interest which are less critical, or irrelevant, are labeled with letters B and C and a number, respectively. In addition, we labelled the whitespace around the critical areas as ZZ.

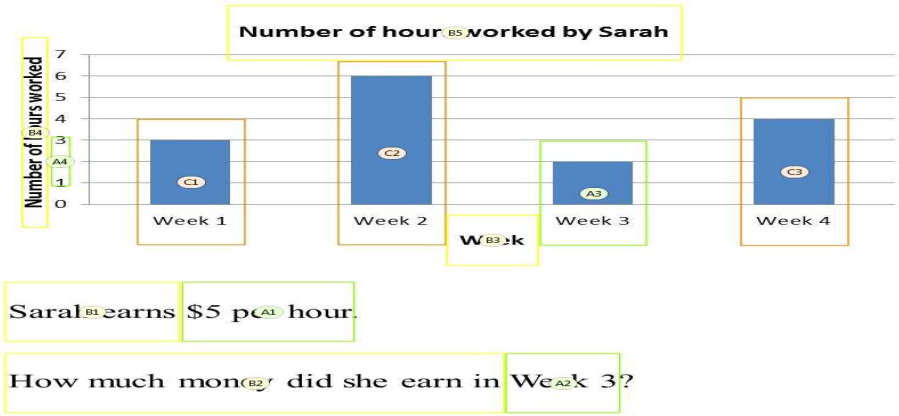

Figure 1: partition of the task sheet into areas of interest (AOI's).

As a result, data items look like the following:

{{227, A1}, {563, B2}, {267, C2}, … , {287, C2}, {517, A1}, {1443, A3}} -> 1,

Figure 2: a typical data item with pairs of elements corresponding to durations in milliseconds (numbers) and AOI's (letter and number juxtaposed) of the fixations, and the result after an arrow.

In the example above, the first fixation occurred on area of interest A1 and lasted 227 ms, the second one on AOI B2, with a duration of 563 ms, and so on. At the end the arrow with a 1 after it indicates that the child solved the problem correctly.

Finding a small and highly representative subset of data (developing a training set).

In order to find small and highly representative sets of data items corresponding to correctly and incorrectly solved instances of the task, we tried to find the smallest subsets of data items (henceforth called *training sets*) on which we could generate classifiers that predicted outcomes with a high degree of accuracy. After a building classifiers based on randomly selected subsets of data items, we could generate a classifier that correctly predicted up to 75% of the test results, and this was using only four data items in the training set (3.5% of the *sample*). It would have been impossible to test all sets of four data items out of 113 (there are 6,438,740 such combinations) so we made a number of classifying testing runs for randomly selected subsets of size 4, and chose some of those sets which yielded classifiers with a high prediction rating. We then inspected the videos of some of these sets and tried to observe what might have been visually outstanding in these. Our prediction rate is marginally better than human experts can do after training on very large data sets. In the world of machine learning, a rating of 75% with a training set of size 3.5% is an extremely good result in what is called supervised learning (since the training set we found is so small, this is called semi-supervised learning (for machine learning principles, consult Hastie, Tibshirani & Friedman, 2009).

From this inspection, we detected parameters to investigate further with machine learning and other techniques, including sequencing, duration and number of fixations and other more elaborate metrics.

Analysis type 1: the order of fixations in the sequence – does it matter or not?

One question we had was whether the order of fixations in the sequence matters, or whether there is something else at work. Some literature in psychology indicates that the order of fixations affects certain cognitive function such as memory (e.g. Bochynska, & Laeng, 2015; Rinaldi, Brugger, Bockisch, Bertolini, Girelli, 2015). First, we tested overall order, building a classifier using the entire sample data. Its predictive rate is over 99% (using this technique we get only one mismatch between predicted and real outputs, due to a faulty item which we were able to locate through the application of the classifier itself). We then permuted the order of the fixation duration and AOI pairs at random in the optical tracks, and passed the permuted input data through the classifier we obtained using the entire sample. Even with the permuted data, we obtain a classification rate which is over 97%. From this we cautiously concluded that the order of the fixations in the sequence has little impact on whether the child responds to the question

accurately.

As an additional check, we investigated whether the order of fixations within critical AOIs mattered. If this were occurring, it might distinguish *understanding* and *interpretation* of the graphical information (Curcio, 2010). In order to study this, we extracted just the pairs of elements corresponding to critical elements, and eliminated the rest of the data elements. With these modified data items, we built a classifier, using training sets of size 13 (approximately 11% of the sample size), and passed the rest of the modified data items through the classifier. This resulted in a prediction rate of up to 66%, which is good but not nearly as good as we had hoped. This indicates that the order in which students inspect critical areas might be of some importance, and it deserves further study. This also led us to a different form of analysis (type 3), even though much more needs to be done than we did here.

Analysis type 2: number of fixations and duration of engagement on task.

The number of fixations and their duration (see figure 3) for the subjects is extremely revealing even though the analysis is less complex. These fixation duration profiles could be interpreted like a simple fingerprint of student engagement and ability. Our analysis of the number of fixations and their duration gives a clear indication that optical tracks can be quite revealing about what the students can or actually do. To state the results briefly, children who respond correctly take a short amount of time (under 30 000 ms) to provide and answer and have a smaller number of fixations (mean of 69) than children who respond incorrectly. Most of the children who respond incorrectly take at least 35 000 ms to respond or have more than 69 fixations. The statistically significant duration averages for children who respond correctly and those who do not are 30 000 ms and 35 000 ms respectively, and 69 fixations vs 77 fixations respectively. Interestingly, a few children (34 out of 113) who take a short amount of time and have a small number of fixations, typical of children with a correct response, provided an incorrect response. In most of these cases children had gathered the correct information from the graph but had made a calculation error. There are 17 children for which we have not yet determined an adequate explanation of their performance. Had those children read the graph incorrectly? Had they understood the task? When interpreting the graph and performing the computation, did concepts become confused? We found that these 17 children completed the task very quickly relative to the other participants, with a mean response time of approximately 25 000 ms. This information leads us to speculate that these children may not have been fully engaged in the task or in some respect confused or wandering. In summary, we can pick out, in each case, the children according to their response from a quantitative analysis by looking just at the duration of their engagement and the number of fixations during their involvement in the task. In the future, we plan to do an Artificial Intelligence based cluster analysis of the number and duration of fixation profiles only, hoping that they will separate out into four categories: those of children who respond correctly, those of children who do not read the graph correctly, those of children who read the graph correctly but miscalculate, and those of children who “do something else”. There is interest and possibly a growing body of work around this topic,

whether it is possible to classify gaze patterns according to the state of mind of the participant subject. It is definitively one of our goals in this and future research (e.g., Horrey, Lesch, Garabet, Simmons, Maikkala, 2017).

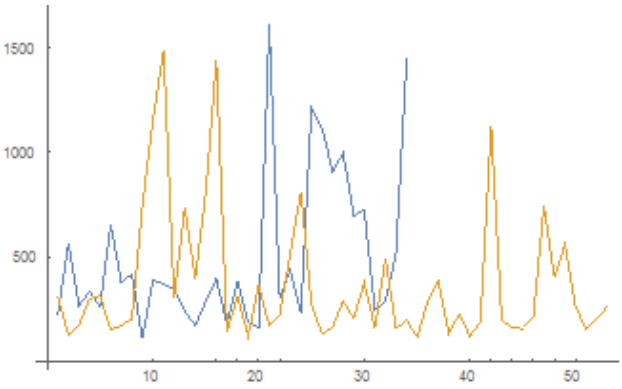

Figure 3: number of fixations and duration profiles of successful child (blue) and unsuccessful child (orange) – the x-axis is the number of fixations, the y-axis is time, the duration of fixations, in ms

Analysis type 3: duration ratios and frequency ratios.

From viewing the videos it appeared that children who get the problem right seem to spend a substantial amount of time looking at critical data, and seem to look at such data more frequently. These parameters were assessed quantitatively, making a distinction between the importance of the area of interest (e.g. A, B, C), and not between the areas themselves (e.g. A1, A2, A3 etc.). Thus, we measured the total amount of time a subject spent looking at critical AOI's (with labels Ax), and non critical areas (Bx, Cx, and ZZ), and also measured the frequency with which a subject inspected an AOI labeled with A, B, C, or ZZ. The total duration of fixations on areas A, B, C, ZZ became DA, DB, DC, and DZZ, and the we considered the ratio DA/(DB+DC+DZZ). We then computed the means of this ratio for the students who successfully solved the problem and for those who did not. The means were used to compute a threshold value and make predictions as to who would successfully solve the problem or not. The same approach was used for frequencies (call the total frequency on A-critical areas FA, FB for B-critical areas, FC for C-critical areas, and FZZ). We computed an analogous ratio where the quantities FA, FB, FC and FZZ were weighted by coefficients 1, .5, .25, and 0, respectively. The rationale for using weights in the case of frequencies is to account for the fact that looking at less critical AOI's, for example, whitespace (ZZ), can easily occur as a result of distraction while inspecting the graph or while moving from a fixation in an important area to another one, and therefore, they are overrepresented and should carry a smaller weight in the frequency count. We acknowledge there are alternative approaches that could be used.

With the two thresholds used in combination one can predict the results with an accuracy of 77%. The thresholds were combined in such a way that if a child spent both, enough time on critical areas, and looked at them frequently enough, the result would be success, and otherwise, it would result in an incorrect response. So it seems that both these parameters are indicative of

a child's ability to successfully solve the graph interpretation task. A post-hoc statistical analysis was done on the means obtained for the duration ratio and the frequency ratio to show that they differ in a statistically significant way. Assuming a normal distribution of the duration ratios, the means of children who were successful and unsuccessful were 1.13 and .76, with a standard deviation of .43 and .42 respectively. These means are statistically significantly different with a p value of 3.32×10^{-36}. Similarly, having tested for the normal distribution of frequency ratios the means are 1.81 and 1.33, with standard deviations of .53 and .54, and a p value of $1.79\times10^{-26,}$, showing again a very significant difference.

A note about validity and reliability.

The results discussed here would need to be validated with further experimentation. For example, do the results hold if the experiments are repeated with systematic variations, changing the height of the bars, the number of the week, and the salary for Sarah? Similarly, do the results remain invariant cross-culturally? We have thought of replicating the experiments, with children of the same age and/or background knowledge, in different English speaking countries and in different cultures with different languages. This work remains to be done. The reliability of these results is given in as much as the calculations are straightforward and easy to check, and the data is clean data as provided by a commercially tested device. It is hoped that in the future, a functional version of the paper can be republished in a way that the reader can verify the programs and use the programs with his/her own data.

Conclusions and direction of future work.

In this report we have discussed the kind of visual processes that might be at work when a child is solving a graph interpretation task, a discussion derived from a machine learning analysis of eye-tracking data collected during the problem solving sessions. It would seem that there is strong evidence to support the claim that the order of the fixations during the problem solving session plays almost no role in the child's ability to succeed in the problem solving task. It would also seem that the amount of time and the number of times spent looking at areas where there is information which is critical for the solution of the problem relative to the amount of time and frequency of glances at other areas is definitively an important indicator of a child's ability to successfully complete the task.

As to how these results would affect teaching practices, one could conclude that it is important that the teacher directs the student attention to what the critical information might be, where it might be located, and how to use it when teaching how to interpret graphs of this sort.

There are many other measures that can be studied (or have been studied, but are not reported here). We mention just a few, without further explanation: string edit analysis, lag analysis, cluster analysis. The limit in how to analyze gaze tracking data is our imagination, in so far as how much information one can glean with computational and mathematical means out of the data in hope of finding useful information.

Acknowledgements.

The first author wishes to acknowledge very valuable on-going discussions with Nora McIntyre, at Psychology in Education Research Centre in the University of York, with whom very valuable discussions ensued about the possible measures discussed here, and about classifying gaze patterns according to the subjects state of mind; and the ongoing support of Prof. Markku S. Hannula at the Dept. of Teacher Education of the University of Helsinki during this research.

References

Australian Curriculum, Assessment and Reporting Authority (2016). *The Australian Curriculum: Mathematics, v8.2*. http://www.australiancurriculum.edu.au/

Bochnkska, A., & Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate the retrieval of visuospatial information. *Cogn Process* (2015) 16 (Suppl 1) 159–163.

Curcio, F. R. (2010). *Developing data-graph comprehension in Grades K-8* (3rd ed.). Reston, VA: The National Council of Teachers of Mathematics.

Garcia Moreno-Esteva, E., Toivanen, M., & Hannula, M. S. (2016). When does visual information become relevant in a dynamic problem solving task in the classroom) – an eye tracking case study. In *International Conference of Mathematics Education Proceeding (ICME-13),* Hamburg, Germany, 2016.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York.

Horrey, W. J., Lesch, M. F., Garabet, A., Simmons, L., Maikkala, R. (2017). Distraction and task engagement: How interesting and boring information impact driving performance and subjective and physiological responses. *Applied Ergonomics*, Volume 58, 342-348.

Lai, M. L., Tsai M.-J., Yang, F.-Y., Hsu, C.-Y., Liu T.-Z., Lee S. W.-Y., Lee M.-H., Chiou, G.-L., Liang, J.C. and Tsai C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review* 10 (2013) 90-115.

Rinaldi, L., Brugger, P., Bockisch, C. J., Bertolini, G., Girelli, L. (2015). Keeping an eye on serial order: Ocular movements bind space and time. *Cognition*, Volume 142, 291-298.

Tobii Technology. (2014). User manual: Tobii TX300 eye tracker, Revision 2. Sweden: Tobii Technology.

White, S., Wood, J., Black, A., & Sampson, G. (2015). Exploring what eye tracking can reveal about student processing of mathematics tasks. In Book of Abstracts of the European Association for Research in Learning and Instruction (EARLI) 30th Meeting (2015), p. 527-528, Limassol, Cyprus.

Wolfram Language and System Documentation Center (2016). http://reference.wolfram.com/language/ref/Classify.html?q=Classify