Application of mathematical and machine learning techniques to analyse eye tracking data enabling better understanding of children's visual cognitive behaviours

Enrique Garcia Moreno-Esteva[a], Sonia L. J. White[b], Joanne M. Wood[c], Alex A. Black[c]

[a]Department of Education, University of Helsinki, Finland
[b]Faculty of Education, Queensland University of Technology, Australia
[c]Faculty of Health, Queensland University of Technology, Australia

Abstract

In this research, we aimed to investigate the visual-cognitive behaviours of a sample of 106 children in Year 3 (8.8 ± 0.3 years) while completing a mathematics bar-graph task. Eye movements were recorded while children completed the task and the patterns of eye movements were explored using machine learning approaches. Two different techniques of machine-learning were used (Bayesian and K-Means) to obtain separate model sequences or average scanpaths for those children who responded either correctly or incorrectly to the graph task. Application of these machine-learning approaches indicated distinct differences in the resulting scanpaths for children who completed the graph task correctly or incorrectly: children who responded correctly accessed information that was mostly categorised as critical, whereas children responding incorrectly did not. There was also evidence that the children who were correct accessed the graph information in a different, more logical order, compared to the children who were incorrect. The visual behaviours aligned with different aspects of graph comprehension, such as initial understanding and orienting to the graph, and later interpretation and use of relevant information on the graph. The findings are discussed in terms of the implications for early mathematics teaching and learning, particularly in the development of graph comprehension, as well as the application of machine learning techniques to investigations of other visual-cognitive behaviours.

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

Corresponding author: Enrique Garcia Moreno-Esteva: PL 9 Siltavuorenpenger 5A, 00014, Finland.
enrique.garciamoreno-esteva@helsinki.fi *DOI: https://doi.org/10.14786/flr.v6i3.365*

1. Introduction

Eye tracking is rapidly becoming an established technique for investigating the cognitive processes involved in learning mathematics and other subjects. The use of eye tracking makes it possible to make implicit visual-cognitive behaviours explicit, in order to better understand learning processes and subsequently inform educational practice (Lai, et al., 2013).

Previous methods, such as interviewing and think aloud protocols, have been adopted to understand the different approaches and strategies used in a range of problem-solving tasks, however, there are limitations to these approaches. First, interviews undertaken following task completion are reliant on a participant accurately remembering and recalling specific steps. Second, think aloud protocols assume the participant has the cognitive flexibility to think out loud while also engaging in the problem-solving task (Rosenzweig, Krawec, & Montague, 2011). For example, to understand how young children engage in mathematics problem-solving tasks, the additional cognitive demand of think aloud protocols could potentially have an adverse impact on task performance. Indeed, Kotsopoulos and Lee (2012) used modified think aloud protocols and real time naturalistic analysis of students completing mathematical problem-solving tasks, and reported that problem-solving often broke down in the early stages of understanding the task. The limitations with many of these approaches led Van Gog, Paas, van Merriënboer and Witte (2005) to highlight the need for research methods to understand different problem-solving processes and supported the use of eye tracking methods for more complex tasks that involve a sequence of cognitive steps.

An extensive body of eye tracking research has focused on the link between visual gaze and information processing. For example, how a student looks at a diagram is influenced by their preliminary intuitions and the conceptions activated by the task context (Knoblich, Ohlsson & Raney, 2001). Eye tracking research has revealed that more experienced problem-solvers (experts) can identify task relevant visual information more rapidly than less experienced individuals (novices), and their visual attention (eye fixation scanpaths) tend to be more focused on relevant than irrelevant regions of the visual stimulus (Gegenfurtner, Lehtinen & Säljö, 2011; Tsai, Hou, Lai, Liu, & Yang, 2012); these objective findings were corroborated by self-reported accounts of participants completing the task (Tsai, et al., 2012). Another study reported that novices display significantly more shifts in visual attention than experts and have longer gaze sequences (or scanpaths) for a given problem-solving task (Kim, Aleven & Dey, 2014). Furthermore, psychology research indicates that the order of fixations affects cognitive functions, such as memory (e.g. Bochynska & Laeng, 2015; Rinaldi, Brugger, Bockisch, Bertolini, Girelli, 2015). Bochynska and Laeng (2015) used eye tracking in a visuospatial memory recognition task and found that both the spatial information and the order of fixations were important for visuospatial memory formation, with increased accuracy in trials where the elements were presented serially, in the same order as in the participant's original fixation scanpath.

The current research, utilised machine learning techniques to make sense of the order structure of the multiple scanpaths of primary school children while they completed a graph problem-solving task, in order to better understand the visual cognitive behaviours involved. The rationale for using machine learning analysis is that it allows us to examine the sequential (temporal) structure of the data. This is unlike the approach adopted in most eye tracking studies and allows novel insight into the underlying behaviours of children while completing these graph tasks. While there are other methods to examine the temporal structure of data, we felt our approach was the most suitable for our research purpose. Our method is also faster and more automated than many other traditional methods. In order to interpret the scanpaths, our approach used the sequential framework of children's data comprehension, described by Curcio (2010) which comprises; understanding, interpretation, and prediction with data. The first level of comprehension, 'understanding', requires reading of the information explicitly stated in the presented data (e.g. graph). The second level of comprehension, 'interpretation', requires reading between the data and integration of the presented information, this level of comprehension requires specific skills, such as comparison or computation (e.g. addition, subtraction, multiplication, division). The third level of comprehension, 'prediction', requires reading beyond the data and the use of existing knowledge to make inferences/predictions from the data. When making predictions, the information is neither explicitly nor implicitly presented in the graph. The bar graph task used in the present study required each child to engage in the first two levels of data comprehension: 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. Most importantly, with

graph interpretation, both visual and cognitive integration is required (Ratwani, Trafton & Boehm-Davis, 2008), where integration of relevant visual attributes on the graph, such as labels, pattern recognition, or other spatial features contribute to higher order visual clusters of information, that are compared to create a coherent representation and response.

The aim of this research was thus to use mathematical and machine learning based analysis of eye tracking data to better understand the visual and cognitive behaviours associated with the completion of a graph task in a sample of children in Year 3. Using a desktop eye tracking system, children completed a mathematics task that involved comprehension of a bar graph. The scanpaths and the accuracy of individual responses to the task were analysed to identify the visual cognitive behaviours when completing the graph task, for example, what happens when children are confronted with such a task, which features children look at, and the order in which relevant information is accessed. The research also aimed to determine whether the gaze patterns for children who correctly or incorrectly completed the task were different, and if so, to identify the characteristics of the different visual cognitive behaviours. Such information will enable more reliable inferences about the implicit visual cognitive behaviours of children when engaging in a graph problem-solving task and other cognitive tasks.

2.   Method

2.1   Participants

Participants included 106 Year 3 children (58 females, 48 males; mean age 8.8 ± 0.3 years) who completed a graphical mathematics task. Children were from three primary schools in South East Queensland, Australia. Data collection occurred in the last half of the school year, and the graph task formed part of a larger study that involved eye tracking while children completed a series of mathematics and reading tasks. All data collection occurred in a quiet room near the respective classrooms. The study was approved by University Human Research Ethics Committee that operates within the Australian National Statement on Ethical Conduct in Human Research. Approval to conduct research in Queensland State Schools was also granted by the Queensland Government, Department of Education and Training.

2.2 Task

The design of the graph task was based on the Year 3 Australian Mathematics Curriculum where children are interpreting and comparing data displays (ACARA, 2016). As presented in Figure 1, the graph task included: a) a bar graph, where the height of each bar indicated the number of hours worked by Sarah during a given week; b) a labelled 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 question related to Sarah's wages in Week 3.

2.3 Apparatus

A screen-based Tobii eye tracker (TX300) operating at 300 Hz recorded the eye movements of the children as they completed the graph task. The task was presented on a 23 inch screen and participants sat comfortably (without restraint) at a working distance of approximately 60cm. The calibration used the TX300 nine point calibration procedure, with re-calibration conducted for any points where calibration was recorded as poor (denoted in red). Only when the calibration procedure was completed for all nine points (denoted in green) was the eye tracking task started. Events were detected using a dispersion based algorithm for detecting fixations. A fixation was defined as static eye movements with gaze positions remaining within a visual angle of 1.6° for at least 100 milliseconds (Tobii Technology, 2014). For the graph task, the 106 participants had a mean tracking percentage of 87.8 ± 9.8%. The output of the eye tracking device is typically a sequence of coordinate pairs relative to the scene that the participant is viewing, together with a time stamp when the locus

of gaze is positioned at a given coordinate pair.

2.4 Data extraction

The visual stimulus (the graph on the screen; Figure 1) was subdivided into regions known as Areas of Interest (AOI). Initially, AOI sequences were generated from the sequence of fixations, as determined by the dispersion-based algorithm. The first step involved determining which were the most critical AOIs and labelling them as A1 (wage information (part of sentence below the graph)), A2 (week number (part of sentence below the graph)), A3 (week 3 bar), and A4 (number region containing the number of hours corresponding to week 3). These AOIs were considered as the minimum number of areas that needed to be fixated to complete the graph task successfully. The other AOIs were labelled as somewhat critical (B) and less critical (C). The categorisation of B areas as somewhat critical was derived as part of an iterative process, incorporating typical classroom practice and qualitative inspection of the scanpaths. Typical classroom practice, or the procedural frameworks used to support children engaging with graphs, usually includes reading the title and axis labels as an initial orientation to the graph; these areas were therefore labelled as somewhat critical B areas. Furthermore, qualitative inspection of scanpaths revealed that when reading the sentences below the graph, some children failed to read the full sentence and did not access the most critical information (A1 and A2), but did read the first part of each of the two sentences. To enable identification of this type of incomplete reading behaviour, the sentences were separated into two different areas, indicating somewhat critical (B) and most critical (A) information. Finally, the less critical C areas were classified because they represented procedural checking that may be promoted as part of classroom practice, for example, systematically checking all bars on the bar graph before providing a response.
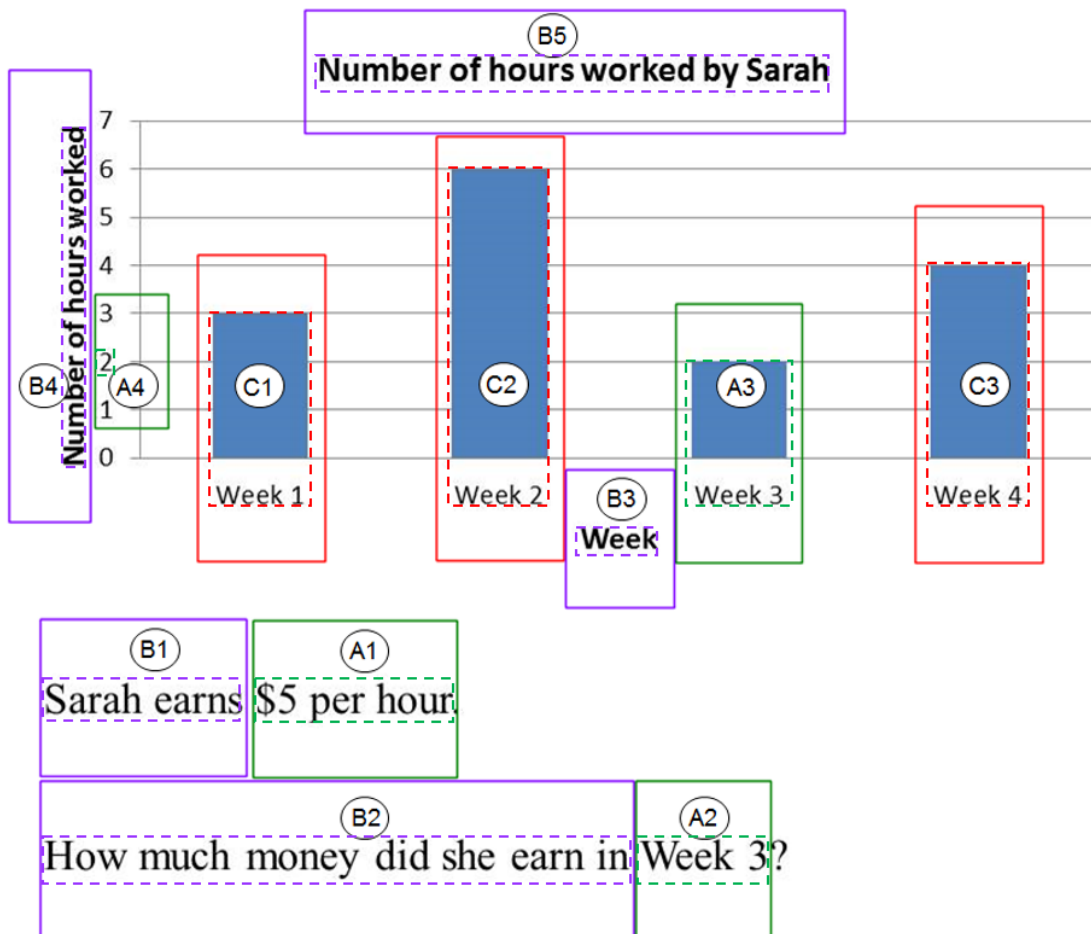

Figure 1. The graph task visual stimulus partitioned into areas of interest (AOIs). In this diagram, dashed lines indicate the original stimulus size and solid lines indicate the alignment of displacement zones for each AOI.

The size of each AOI regardless of whether they were categorised as A, B or C, was calculated using two criteria; first, the AOI had to allow for vertical and horizontal displacement errors in the visual scanpath, and second, the AOIs could not overlap (Holmqvist, et al., 2011). While some studies involving children (e.g. Sasson & Elison, 2012) have implemented a two degree vertical and horizontal displacement zone centred over relevant visual stimuli (one degree above and one degree below the stimulus), this was not possible with the current graph task as it would have resulted in substantial overlap of AOIs. Thus wherever possible, a 1.9 degree vertical and 0.5 degree horizontal displacement zone was centred over the stimulus and in the event of overlap, the horizontal displacement zone was reduced to 0.15 degrees, which occurred for the following AOIs: A1 and B1, A2 and B2. To avoid overlap between C2 and B5 AOIs, B5 had a 1.5 degree centred vertical displacement, whereas the C2 vertical displacement zone could not be centred over the stimulus, so the C2 AOI was defined as 0.5 degree above and 0.95 degree below the stimulus. To distinguish fixations on the y-axis, A4 and B4 had a 1.9 degree vertical displacement, and a larger 1 degree horizontal displacement.

The data sequences that were obtained initially were based on fixations and then transformed into dwell sequences by coalescing contiguous identical elements in the sequence. Thus, for example, if in a fixation sequence we have B2, A1, A1, C3, this would give rise to the dwell sequence B2, A1, C3. In our analysis, sequences of dwells were used. The data included 106 data sequences or scanpaths (known as the inputs - both terminologies are used to reflect what is typically used in the literature), one for each child, and the corresponding 106 answers to the graph question (the outputs), were categorised as correct (1) or incorrect (0). The data sequences (sequences of AOI names) could thus be grouped into two classes, one being the sequences of children who responded correctly to the task, and the other being the sequences of children who responded incorrectly.

2.5 Machine learning techniques

Machine learning approaches were used to find a representative sequence of AOIs that would provide qualitative information about the visual cognitive characteristics associated with task performance. Two methods are described that provide means by which an average scanpath can be extracted from a set of multiple sequences or scanpaths obtained from children performing the same task. Method one applies a naïve Bayesian approach to calculate the most probable vector for each class (using two possible alternatives to managing variable feature vector length), and method two applies a single iteration of a K-means approach to calculate the central vector for each class using an edit distance metric. The use of these techniques will form the basis for future research that characterises the relevant scanpaths of children to potentially guide identification of their ability to effectively perform graph tasks.

2.5.1 Method one – most probable vector for each class

The naïve Bayes model or classifier "learns" from the data sequences, which are known as feature vectors, which is a commonly used term in the machine learning literature (Murphy, 2012). A data sequence can be defined as a finite list of representational feature objects or values, which can be names, numbers or symbols, as long as the possible set of representational names or values is finite. In the case of scanpaths, the feature values are the AOI names. The classes in the model are the children who either correctly or incorrectly completed the graph task.

Learning a naïve Bayes model consists of calculating the class conditional probabilities of the feature values. For each class and each feature, the number of occurrences of the feature value is divided by the number of feature vectors in the class (whether the child completed the task correctly or incorrectly) in question. This yields the class conditional probability for the possible values of a given feature in the given class. The "classifier" obtained from the model consists of probability distributions of the feature values determined by the class conditional probabilities for the values of each feature in each class. For example, suppose that we have feature vectors (each one in brackets) in a class: (A, **B**, C, A, B), (B, **B**, C, C, B), and (C, **A**, B, C, A). Then, the class conditional probabilities for the second feature value (in bold) of this class are, for feature value A, 1/3; for feature value B, 2/3; and for feature value C, 0 because it does not appear in the second feature position.

For a given sequence or feature vector in a given class, all the class conditional probabilities are multiplied and the product is multiplied by the class prior probability. In our analysis we assumed what is called a uniform prior, that is, a prior probability value of ½ for each class. The final products, one for each class, are the joint probabilities. To obtain the most probable feature vector for a class (i.e., the most probable or average scanpath or data sequence for the given class) the feature value that has the highest probability of occurring in a feature and class, as given by the class conditional probabilities, is selected for each feature and class. In the above example, the resulting vector would be (A, B, C, C, B). In the first position, all values have probability 1/3 so we pick one at random, say A. In the second through fifth positions, B, C, C, and B have probability 2/3 whereas the other symbols in each position have probability 1/3 or 0. The resulting list, or vector, of most probable feature values is the most probable feature vector (or average scanpath or data sequence) for that class.

The vectors obtained for each class provide qualitative information regarding the children's task performance according to the previously established criterion, which indicates whether the graph task was completed correctly or incorrectly. The feature vectors can have different lengths, which can lead to biases; there are two alternatives for managing variable feature vector lengths.

The first way to deal with this length variability is to consider those feature vectors in a class which are shorter than the longest one in that class, as feature vectors containing missing data. The class conditional probabilities are calculated as outlined in the example. If a feature is considered as a coordinate or location in a vector, for a given feature and class, the probability of a feature value is the number of times that value occurs in any of the feature vectors, at the given location, divided by the total number of feature vectors in the given class. If a feature vector is too short to contain any feature values, then it will not contribute to the numerator of the preceding quantity, as in the example above.

The alternative approach to handling the length variability is to turn all the feature vectors in a given class into vectors of the same length by "padding" the shorter vectors in the class with a dummy value to the right, up to the length of the longest feature vector in the class. Padding a short vector to the right means that, after the last value of the short feature vector, a new padded value is added to the end of the list of feature values until it is the same length as the longest feature vector in the class. The "padding feature value" instances are different to all original feature values occurring in the data feature vectors. As an example, if the longest feature vector has a length of 5 (A, B, A, C, A), and we have a feature vector of a length of 3 (A, B, C), then, in positions 4 and 5 of the short feature vector we introduce the padding feature value, say X, so that instead of (A, B, C) we end up with (A, B, C, X, X). The class conditional probabilities are calculated as before, noting that in some cases, the most probable feature value might turn out to be the padding feature value.

In the case of our analysis we used the first method to deal with length variability because, among other more technical reasons, it allowed us to have strings that contain only AOI names.

2.5.2 Method two – the central vector for each class

The second overall method of analysis involves taking a measure of the "distance" between one feature vector and another one with the Levenshtein edit distance metric. The edit distance between two feature vectors is given as the number of feature deletions, insertions, and replacements that are required to transform one of the vectors into the other one. The method then works by finding a feature vector, which is called a central feature vector, that minimises the total edit distance between the central feature vector and each of the other feature vectors. Methods to compute edit distance and thus a central feature vector are well understood in the computer science or mathematics literature (see Skiena, 2010 for a review). These central feature vectors also reflect, on average, what is happening with all the sequences, and should be generally similar to those obtained with method one - the most probable vector.

3. Results

The results of method one show that classifying multiple scanpaths, particularly, dwell area of interest (DAOI) sequences with naïve Bayes works extremely well (where a dwell is a continuous series of fixations within the same AOI). This in itself is informative, as it means that the elements of a DAOI sequence are not necessarily correlated, that is, after looking at any one particular AOI, the observer may look at any other AOI with equal likelihood. Both method one and method two allowed us to produce "virtual scanpaths" (the most probable vector and the central vector scanpaths) that demonstrate what children are doing when they are solving the problem correctly or incorrectly.

3.1 Using a naïve Bayes classifier to obtain most probable vectors

Using a naïve Bayes classifier we were able to get a perfect classification, where our training error rate was zero. Since predictive analyses was not our goal, we did not write our own cross validation routine. However, with a commercial package that handles missing data very differently (Mathematica, Wolfram Research), we obtained a training error rate of .15 and a "leave one out" cross validation error rate of .27. This suggested that the AOIs in the gaze sequences are uncorrelated along the sequences and obey the naïve Bayes assumption. What makes the use of a naïve Bayes classifier attractive to us is the possibility of generating "virtual scanpaths" that are "most probable". Since for each feature (each coordinate of the scanpath sequences) and for each class we obtain a probability distribution of the possible AOIs, the most probable one can be selected. This provides us with an exemplary sequence that provides good qualitative data regarding what the children are doing while they solve the task, depending on whether they do so correctly or incorrectly. On average, the lengths of the scanpaths of children in either group are largely similar (69 ± 31 AOIs for children who responded correctly, vs 77 ± 39 for children who responded incorrectly), although they are slightly shorter for children who completed the task correctly. The qualitative information provided by the average scanpaths becomes even more evident with further manipulation and post-processing of the average scanpaths. If we merge the AOIs which are contiguous and identical in the average scanpath into a single instance, and remove dwells on blank spaces, the average scanpaths become much shorter, and this varies between the two classes. The correct children's post processed (merged) average scanpath is as follows:

{C1,B1,A1,B2,A2,B2,A2,A3,B2,A3,B2,A3,A1,A3,A1,A3,A1,A2,A1,A3,A1,A3,A1}

And the incorrect children's post processed (merged) average scanpath is as follows:

{C2,C1,B1,A1,B1,A1,B1,B2,B1,B2,A1,B2,C2,B2,A1,C2,B2,A3,B2,C2,A3,A1,A3,A1,C1,C2,B2,A2,A3,B2,A1,B2,A1,A3,C2,B3,C2,B2,A3,A2}

As can be observed, the incorrect children's merged average scanpath is almost twice as long (40 AOIs) as that of correct children (23 AOIs). It is notable that A4 does not feature in the merged average scanpath for either the correct or incorrect children. This is likely to indicate that while children may have accessed A4 in their scanpath, there is not one clear position in the participants' sequences where A4 features, therefore it does not appear in the average scanpath. The percentage of overall AOI accessed in the merged average scanpath are presented in Figure 2. This shows that the gaze of the incorrect children wandered more than the gaze of correct children. It is also interesting to note that the incorrect children spread their attention relatively evenly across the A, B, and C areas, whereas the correct children exhibited greater visual attention (i.e. highest percentage of dwells) on the more critical A areas.
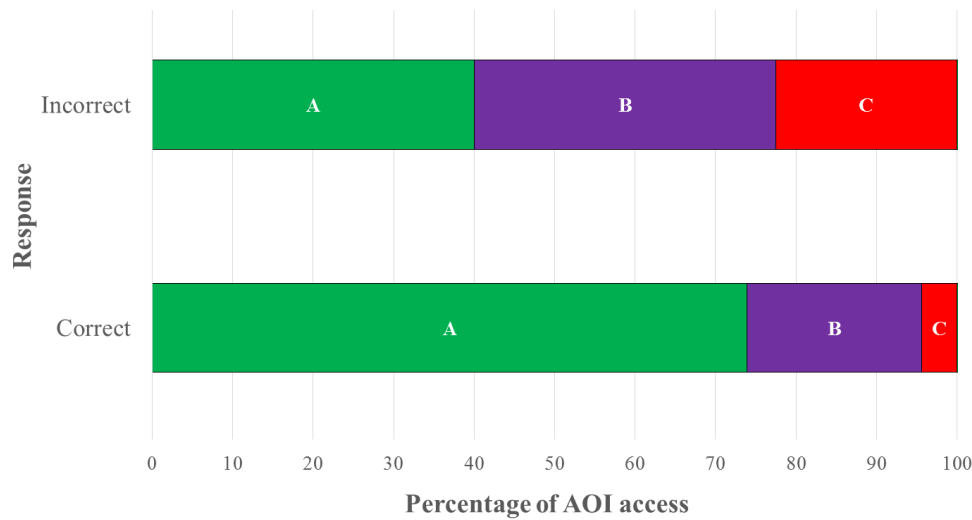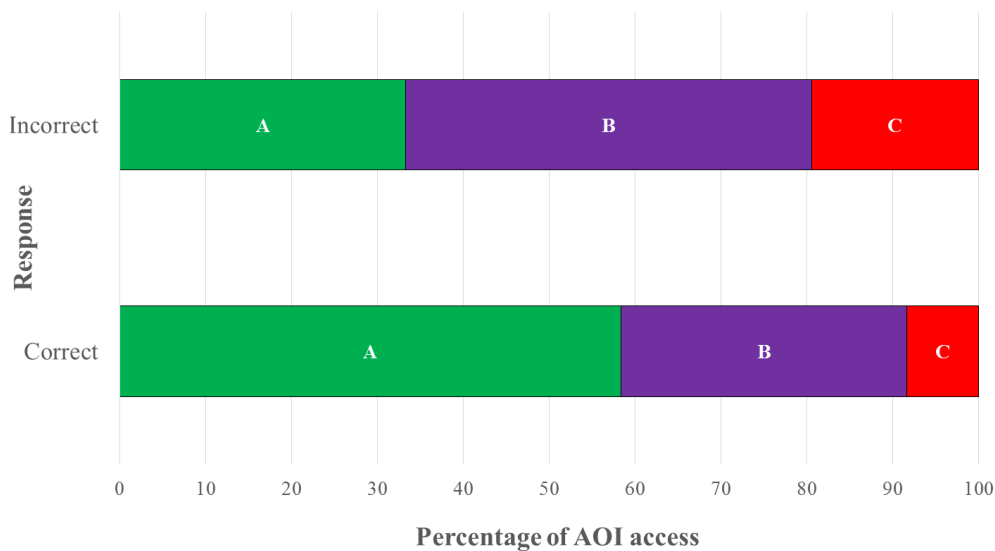
Figure 2. Percentage of AOI access in the merged average scanpath, comparing correct and incorrect children. A AOIs: most critical, B AOIs: somewhat critical, C AOIs: less critical.

Examining the merged average scanpath as a sequence of dwells in areas A, B and C, demonstrates a more detailed characterisation of the visual cognitive behaviours (Figure 3). Correct children more rapidly accessed and maintained attention on the A areas. Whereas children who were incorrect did not exhibit the same rate of attention on A areas, but rather appeared to be uncertain as to where to focus their visual attention, with many shifts between A, B and C areas with a larger number of dwells.
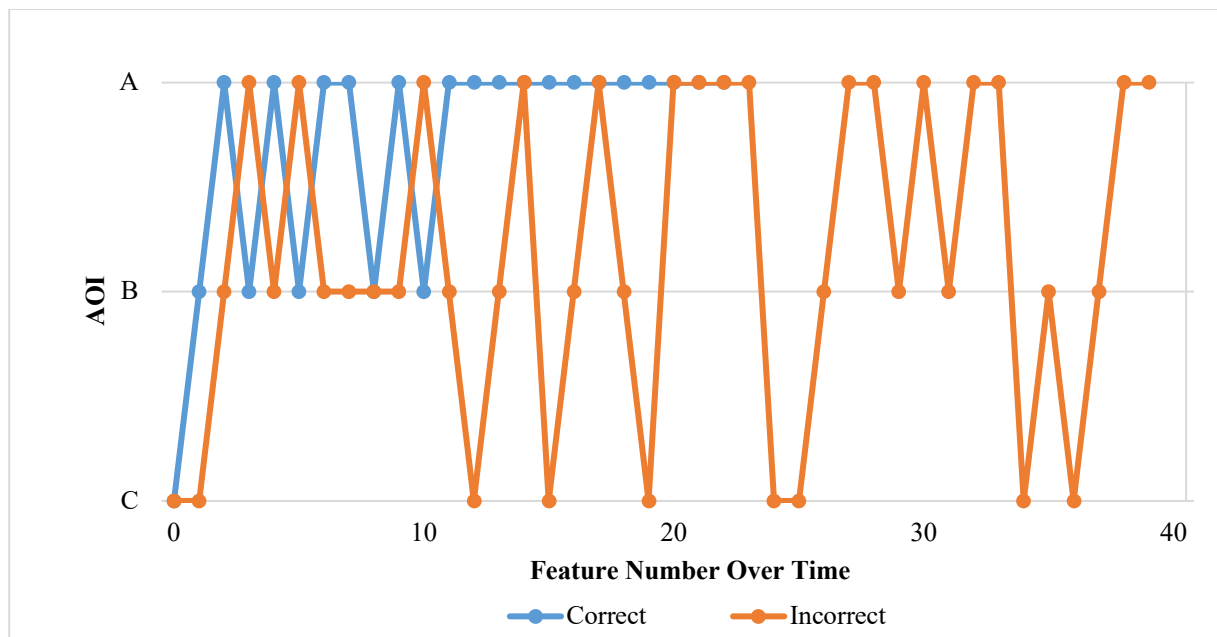
Figure 3. Merged average scanpath, showing which AOI were accessed by correct and incorrect children over time.

3.2 Finding the central data item in each class and what it reveals

The average scanpath derived by method two (central data item) determined that correct children had the following sequence:

{A1,B5,B1,B1,B1,A1,A1,A1,C2,A1,B2,B2,B2,B2,A2,B2,A2,A2,A1,B3,C2,C1,A3,A3,A1,A3,B3,A4,A3,A4,A3,A3,B2,A2,A1,A1}

while the sequence for incorrect children was as follows:

{C1,B1,B1,A1,A1,B2,B2,B2,B2,B2,B2,B2,A3,C1,C2,B3,A3,A4,C1,B5,B2,A2,B2,B2,B2,B2,A2,C3,C1,A4,C2,A3,B1, A1,A2,A1}

The average scanpath sequences for correct and incorrect children have the same sequence length (36 values). In terms of percentage of dwells in the most critical areas, incorrect children, in contrast to the correct children, spent a lot of time looking at the less critical areas (B and C) (Figure 4), which is similar to the results of method one (the most probable vector).

Figure 4. Percentage of AOI access using the central feature vector, comparing correct and incorrect children. A AOIs: most critical, B AOIs: somewhat critical, C AOIs: less critical.

The central feature vectors allowed us to learn how the critical areas were accessed. If we remove all of the non-critical (B and C) areas, leaving only the A areas, a distinction between the two vectors which is crucial to our understanding of the children's task behavior becomes evident. Following this step, we obtained the following correct sequence:

{{A1}, ■, ■, ■, ■, {A1}, {A1}, {A1}, ■, {A1}, ■, ■, ■, ■, {A2}, ■, {A2}, {A2}, {A1}, ■, ■, ■, {A3}, {A3}, {A1}, {A3}, ■, {A4}, {A3}, {A4}, {A3}, {A3}, ■, {A2}, {A1}, {A1}}

And incorrect

{■, ■, ■, {A1}, {A1}, ■, ■, ■, ■, ■, ■, ■, {A3}, ■, ■, ■, {A3}, {A4}, ■, ■, ■, {A2}, ■, ■, ■, ■, {A2}, ■, ■, {A4}, ■, {A3}, ■, {A1}, {A2}, {A1}}

or more simply, without the non-critical areas: Correct response sequence:
{A1,A1,A1,A1,A1,A2,A2,A2,A1,A3,A3,A1,A3,A4,A3,A4,A3,A3,A2,A1,A1}

And, the incorrect response sequence:
{A1,A1,A3,A3,A4,A2,A2,A4,A3,A1,A2,A1}.

Figure 5 summarises the sequence of dwells on most critical areas (A1-A4) and other non-critical areas (B-C) for the modified sequences. Looking at the modified sequences, it is evident that area A4, which includes the values along the y-axis of the task graph, is infrequently visited by incorrect children, and, as anybody who learns math as a student or as a professional is aware, knowing what is being counted is critical to understanding and solving graph problems correctly. Moreover, the correct children's inspection of the most critical areas is in a predictable order (A1, A2, A3, A4). Conversely, a less predictable order was evident in the modified sequence of the incorrect children.
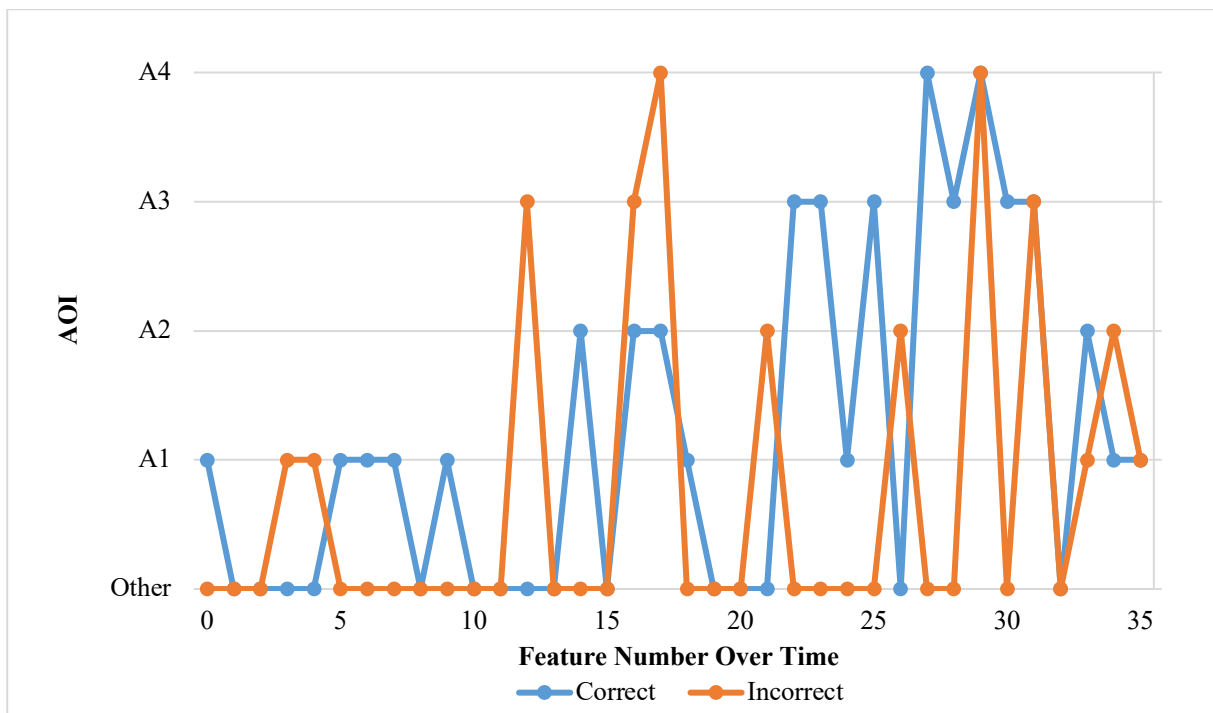
Figure 5. Modified average scanpath, showing the sequence in which the most critical AOIs (A1-A4) and other non-critical areas (B-C) were accessed by correct and incorrect children over time.

4.   Discussion

In this paper we discuss two novel methods of analysing eye tracking data to better understand the visual and cognitive behaviours associated with completion of a graph task in a sample of children in Year 3. Each method produced two strings (one for the correct children and one for the incorrect) which were compared. The initial findings are significant because they demonstrate how the methods developed by our research group can assist in characterising the visual behaviour of children who correctly or incorrectly answered the graph task. The results of both methods of analysis reveal differences in gaze patterns between the correct and incorrect groups. This discussion will summarise the key findings for each method, compare the different analytic approaches and characterise the visual cognitive behaviours identified.

Method one used a naïve Bayes classifier to obtain most probable vectors. The perfect classification training rate with our software of 100% suggests that the eye tracking data for this task obeyed the naïve Bayes assumption, and that the AOIs in the scanpath were uncorrelated, and an average scanpath can plausibly be determined. We believe that this is likely to be task dependent, however, further research is required in order to fully understand this issue. We did not do cross-validation with our software, but used a commercial package (Mathematica, Wolfram Research) which showed a training error rate of .15 and a leave one out cross validation of .27, which given the small amount data and relatively large number of features, is still very good. We are optimistic that with more data, this rate would drop.

With method one, comparisons of the average number of dwells in the sequences of each class identified a slightly higher number of dwells for the incorrect children, compared to the correct children (77 vs 69 AOIs, respectively). This higher number of dwells for the incorrect children is in accord with the findings of Kim et al. (2014) who found that novices display significantly more shifts in visual attention than experts and have longer gaze sequences (or scanpaths) for a given problem-solving task, than their more experienced counterparts. In our study, the correct children exhibited fewer dwells and fewer shifts in visual attention than

their counterparts who were incorrect. This is further evidenced in the merged average scanpath, where correct children had a higher percentage of dwells in the most critical A areas, whereas the incorrect children had dwells distributed across the most critical and less critical (A-C) areas. Interestingly, these patterns reflect those of Gegenfurtner, et al., (2011) and Tsai, et al., (2012) regarding the behaviours of more experienced problem solvers who were shown to identify task relevant features of visual information more rapidly than less experienced individuals, and their visual attention focused more on relevant than irrelevant regions of the visual stimulus. Our research was also able to characterise the sequence of rapid dwells on relevant/critical areas in the children who completed the graph task correctly, as shown in Figure 3.

Method two used the central data item to determine the average scanpath, as well as a procedure to isolate the sequence of dwells on the most critical A areas. This isolation of the sequence of dwells on A areas was revealing in terms of understanding whether there was a logical sequence in accessing the most critical information. The children who responded correctly followed a logical sequence of dwells (based on our initial categorisation of which areas were most or least critical), that progresses from A1 to A2 to A3, then A4 (Figure 5). Conversely, the children who responded incorrectly, accessed A2 relatively late in the sequence. A2 represented the area surrounding the main question 'How much did she earn in Week 3?' with A2 specifically representing 'Week 3'. Delays in accessing this information may result in poorly focused attention to the relevant aspects of the graph, resulting in the child not knowing which bar they should be using in their calculation. Relating this to Curcio's (2010) framework it is the first level in the data comprehension that is delayed, or out of sequence for the children who responded incorrectly. The initial understanding and reading of information in a logical sequence is the essential foundation to guide subsequent visual attention for interpreting and integrating the relevant graphical information and successfully completing the task. This finding is consistent with Kotsopoulos and Lee (2012), who reported that student problem-solving most often broke down in the early stages of understanding a mathematics problem-solving task. It may be that if initial understanding or orientation is not achieved early, the visual dwells become more frequent and variable in location, where children may be looking for information to help them understand the requirements of the task. It may also be that the positioning of the question, either above or below the graph, would impact on the visual behaviours and task response.

In comparing the results of method one and method two, there are similarities and differences. Both methods produce an average scanpath where the correct children have a larger percentage access of A areas, whereas incorrect children distribute their dwells across A, B and C areas, demonstrating that they may not be identifying the most critical areas. The average scanpath derived using the most probable vector indicated that children who provided correct responses had a slightly shorter average scanpath (69 AOIs) compared to the children who were incorrect (77 AOIs). This distinction was not evident in the average scanpath derived using the central data item, both groups had an average scanpath of 36 AOIs. A further difference between the two methods was associated with dwells on A1-A4 AOIs. The average scanpath (central vector) included dwells on all of the most critical areas (A1-A4) and implied a logical dwell sequence, whereas the average scanpath (most probable vector) did not include an A4 dwell. As noted in the results for method one (most probable vector), the omission of A4 from the average scanpath could be an indicator of variation in how children accessed this scale information, so it did not appear using the most probable vector method. For example, children may have counted up the gridlines within the body of the graph to determine the two hours worked in Week 3. Or alternatively, children may have looked more generally at the scale of the y-axis, but not necessarily specifically at A4. The source of this variation would need to be investigated with eye tracking data from another task.

There are limitations to both methods and in the characterisation of the average scanpath for children who responded correctly and incorrectly. As with all averaging procedures, there is some loss of individual detail, however, the aim of this study was to characterise the visual cognitive behaviours of those children who completed the task correctly versus those children who were incorrect. Method one, using a naïve Bayes classifier to determine the most probable vector, might be limited by the data used in the analysis unless, as was the case in our study, the eye tracking data obeys the naïve Bayes assumption of an uncorrelated AOI sequence. Future research should replicate this task, or a close variant, in order to obtain test data to validate the models produced here.

Overall, this research has applied novel machine learning and mathematical analyses to characterise the visual cognitive behaviours of Year 3 children engaged in a mathematics graph task. The resulting

characterisations support the importance of initial understanding of the presented task and identification of the most critical information. While identification of the most critical information may be part of a typical classroom practice, this reinforcement, and the logical sequence of visual access of the most critical information, may be beneficial for children who feel less confident.

Keypoints

For Year 3 children undertaking a graph analysis task:

- Machine learning techniques take the analysis of eye tracking data to a new level. In particular, they allow for understanding of phenomena in the dimension of time in relation to the order structure of data.
- Spending time looking at critical areas and accessing critical areas in a logical order is important for completing the task successfully.
- Spending too much time looking at less critical areas, and more dwells, is indicative of probable failure to successfully complete the task.
- For teaching purposes, it is important to identify critical areas and to break down the process of problem-solving into recognisable steps which can be carried out procedurally.

Acknowledgments

References

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

Bochynska, A., & Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate the retrieval of visuospatial information. *Cognitive Processing, 16* (*Suppl.* 1) 159–163. doi:10.1007/s10339-015-0690-0

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

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552. doi:10.1007/s10648-011-9174-7

Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van De Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. New York, NY: Oxford University Press.

Kim, S., Aleven, V., & Dey, A. K. (2014, April). *Understanding expert-novice differences in geometry problem solving tasks: a sensor-based approach*. Paper presented at the CHI'14 Extended Abstracts

on Human Factors in Computing Systems, Toronto, Ontario, Canada. doi:10.1145/2559206.2581248

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29*(7), 1000-1009. doi:10.3758/BF03195762

Kotsopoulos, D., & Lee, J. (2012). A naturalistic study of executive function and mathematical problem-solving. *The Journal of Mathematical Behavior, 31*, 196-208 doi:10.1016/j.jmathb.2011.12.005

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,* 90-115. doi:10.1016/j.edurev.2013.10.001

Murphy, K. P. (2012*). Machine learning: A probabilistic perspective*, Cambridge: MIT Press.

Ratwani, R, M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied, 14*, 36-49. doi: 10.1037/1076-898X.14.1.36

Rinaldi, L., Brugger, P., Bockisch, C. J., Bertolini, G., & Girelli, L. (2015). Keeping an eye on serial order: Ocular movement bind space and time*. Cognition, 142*, 291-298. doi:10.1016/j.cognition.2015.05.022

Rosenzweig, C., Krawec, J., & Montague, M. (2011). Metacognitive strategy use of eighth-grade students with and without learning disabilities during mathematical problem solving: A think-aloud analysis. *Journal of Learning Disabilities*, *44*, 508-520. doi: 10.1177/0022219410378445

Sasson, N. J., & Elison, J. T. (2012). Eye tracking young children with Autism. *Journal of Visualized Experiments, 61*, e3675. doi: 10.3791/3675

Skiena, S. (2010). *The algorithm design manual* (2nd edition), New York: Springer Science Business Media

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

Tsai,M.-J., Hou, H.-T., Lai, M.-L., Liu, W.-Y., & Yang, F.-Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education, 58*(1), 375-385. doi: 10.1016/j.compedu.2011.07.012

van Gog, T., Paas, F., van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237-244. doi:10.1037/