Scanpaths distinguish problem format in a math cognition task

Samantha Stranc (sam.stranc@carleton.ca)
Institute of Cognitive Science, Carleton University, Canada

Shawn Tan (shawn.tan@carleton.ca)
Institute of Cognitive Science, Carleton University, Canada

Kasia Muldner (kasia.muldner@carleton.ca)
Institute of Cognitive Science, Carleton University, Canada

Abstract

Eye tracking data can inform on cognitive processes. To date, the most common type of analysis corresponds to fixation data. Consequently, less is known about the potential utility of *scanpaths*, which are sequences of eye fixations. In the present paper, we analyzed scanpaths collected as participants solved basic arithmetic problems in two formats: a multiplication format and a division format. The results show that scanpaths do distinguish between the two formats, as reflected by varying similarity scores obtained through the MultiMatch scanpath tool.

Keywords: scanpaths; eye tracking; mental arithmetic; strategies

Introduction

An objective measure informing on cognitive processes corresponds to visual attention measured by an eye tracker. Eye tracking data is commonly analyzed by considering *fixations*, which occur when the eye pauses over a region of interest. While informative, this measure ignores that fixations occur over time in a particular sequence. A sequence of fixations is called a *scanpath.* Recently, the availability of algorithms and tools for scanpath analysis have made this method more accessible, opening the door for research on what insight scanpaths add about cognitive processes over and beyond standard fixation analysis. Here, we add to this work by applying scanpath methods to data from a math cognition task, in order to analyze if scanpaths are affected by different problem formats.

Background: Scanpaths, and Cognitive Processing

To date, eye tracking analysis has predominantly relied on fixation data (e.g., Sharafi et al., 2015, Lai et al., 2013, Mayer, 2010) - for a comprehensive review, see Henderson & Ferreira (2004). Fixation data is popular because (1) it is simple to collect, as eye trackers include built-in algorithms for capturing it, (2) it contains informative features such as the location and duration of visual attention that can inform on cognitive processing, and (3) it is relatively straightforward to analyze (e.g., by comparing fixation counts between conditions). To illustrate some work with fixation data, Susac et al. (2014) reported a negative correlation between number of fixations and problem-solving expertise, indicating that as expertise increases, the number of fixations decreases. Other studies have also found this relationship, in diverse domains like chess (Reingold et al., 2001), epilepsy diagnosis (Balslev et al., 2012), and computer program construction (Nivala et al., 2016).

While fixation data is informative, detecting certain cognitive processes may require sensitivity to not only *where* fixations occur but also the *order* in which they occur. For instance, the order of fixations can inform on viewing patterns. Holmqvist et al. (2011) analyzed viewing patterns for multiple-choice questions on math problems, including overview and focused scanning. *Overview* scanning involved short fixations over the whole problem, while *focused* scanning involved longer fixations in a specific problem region. Since identifying the type of scanning requires more than fixation data alone, the analysis involved data on sequences of fixations, i.e., *scanpaths*. Holmqvist et al. (2011) found that high-ability students had significantly more focused scanning patterns as compared to low-ability students.

A popular application for scanpath data pertains to image and scene analysis. For instance, Foulsham and Underwood (2008) examined if prior experience with a scene would impact scanpaths or if salience was the only predictor of where people would look. The results show that saliency alone is not a sufficient predictor of viewing patterns and that prior experience with a scene influences viewing behavior. Moreover, the results highlight a discrepancy between human-generated scanpaths and ones artificially generated from the saliency map theory, leading the authors to conclude that theories need to be supplemented with sequential aspects of oculomotor control. Coutrot et al. (2018) recorded scanpaths from individuals looking at scenes in three contexts (free viewing, saliency search, and cued object search). The scanpaths were used as input to hidden Markov models, which were subsequently fed to classifiers that predicted the type of viewing context. The results demonstrate that scanpaths can be used to distinguish the type of context with reasonable accuracy.

In addition to scene analysis, scanpaths have been used in other contexts to distinguish experimental interventions and/or tasks. For instance, Zhou et al. (2016) analyzed scanpaths related to different decision-making tasks and conditions (e.g., one task involved choosing between risky options under two different conditions). The similarity of scanpaths in a given decision condition was more similar than between the conditions. This indicates that attentional patterns, and possibly cognitive strategies, were affected by the intervention related to the decision-making condition.

As a final example, research shows scanpaths can distinguish populations in a variety of cognitive tasks. For instance, French et al. (2016) found that the scanpaths produced while solving analogy problems identified with reasonable accuracy whether the participant was an adult or child. As a second example, Von der Malsburg et al. (2017) showed that in a reading task, older readers produced more inconsistent scanpaths compared to those produced by younger readers.

While there may be benefits of taking into account the additional information afforded by scanpaths, a challenge relates to the analysis of scanpaths.

Scanpath Analysis A common approach to scanpath analysis involves quantifying the similarity between pairs of scanpaths. Early methods, such as 'Mannan Linear Distance' (Mannan et al., 1995), provided a measure of scanpath similarity by calculating the absolute distance between the scanpaths' fixation coordinates. This approach largely ignored the order of fixations, a shortcoming that was addressed by other methods, such as Levenshtein string edit (Levenshtein, 1966) and ScanMatch (Cristino et al., 2010). Both of these methods involved the use of 'areas of interest' (AOIs) on the target viewing area. Fixations were labelled by the AOI they appeared in and scanpaths were represented as strings of AOIs. This facilitated scanpath comparison, as similarity between two scanpaths could be measured by the minimal number of changes needed to render the two sequences identical. While ScanMatch improved the string edit method, both methods lack the ability to discern scanpath shape and used only a single measure to characterize scanpath similarity.

In contrast to using AOI's for fixation markers, the MultiMatch analysis tool (Jarodska et al., 2010; Dewhurst et al., 2012) represents scanpaths as a series of geometric vectors, allowing for comparison across five vector dimensions: shape, direction, length, position, and duration. For each dimension, a similarity score ranging from 0 to 1 is produced, where 1 indicates two scanpaths are identical and 0 indicates no similarity between the scanpaths. For the record, we indicate how the dimensions are computed for each feature: *shape* is the difference in saccade vectors $\boldsymbol{u}_i - \boldsymbol{v}_j$, *direction* is the difference in angle between saccade vectors, *length* is the difference in amplitude of saccade vectors $||\boldsymbol{u}_i - \boldsymbol{v}_j||$, *position* is the distance between fixations, and *duration* is the difference in duration between fixations. Recently, analysis comparing MultiMatch and Scanmatch reported advantages for MultiMatch (Dewhurst et al., 2012; Foerster & Schneider, 2013; Gurtner et al., 2019).

Summary As the description above highlights, scanpath analysis has been applied in a variety of contexts, such as reading, analogy making, and decision making. However, to date there is still relatively little work on scanpaths and so additional research is needed to determine the utility of scanpath data.

Present Study

In the present study, we investigated what scanpath data adds to standard eye tracking measures corresponding to fixations. We used data from a prior study (Tan, Muldner and LeFevre, 2016) that we now describe to provide context for the present analysis.

Problem Format and Previous Results

The prior study by Tan et al. (2016) used basic arithmetic problems to explore the impact of two problem formats on solution latency and visual attention. In the *traditional* format, participants were presented with standard division problems (see Figure 1, rows a + b). In earlier work, LeFevre and Morris (1999) found that when these problems involved large dividends, participants reported first converting the problem into a multiplication format and then solving the recasted problem. This *recasted* format was the second type of format used in Tan et al. (2016), where the problem was already formulated in a multiplication format (see Figure 1, rows c + d).

The study involved 33 participants who solved a total of 144 problems (72 in the traditional format and 72 in the recasted format). For each format, factors that were controlled for included: (1) the position of the missing element (either in the 3rd or the 5th position in the equation for each format type, see Figure 1) and (2) operand size, with so-called *small* problems containing dividends smaller than 25 and *large* problems containing dividends equal to or greater than 25 (note: dividend is the first number in the equation for all formats, see Figure 1). The problems were randomly shuffled prior to being presented – details are in Tan et al. (2016). The problems were shown on a computer screen, one problem per screen, and participants were asked to state their response verbally for each problem. The time taken to solve each problem was measured from the onset of the problem to the onset of the verbal response. Once a response was recorded the current problem disappeared and the program moved on to the next problem. An eye tracker (SR Eyelink 1000) recorded participants' visual attention during the experiment. The experiment lasted approximately 50 minutes. After the experiment, AOIs were created around each of the symbols and operators in preparation for analysis of visual attention (AOIs not shown in Figure 1).

The goal of the Tan et al. (2016) study was to confirm prior proposals that participants engage in mental recasting, i.e., when presented with a traditional division-format problem that is facilitated by recasting, do they transform it to a multiplication format prior to solving it? The key method used to answer this question corresponded to analysis of where on the problem elements participants were fixating, as well as qualitative analysis of movements of those fixations over time. The results showed that fixations were affected by problem format. Recasted problems resulted in increased fixation time on the middle element in the problem equation

	Traditional division format				
(a)	72	÷	[ ]	=	9
(b)	72	÷	9	=	[ ]
	Recasted multiplication format				
(c)	72	=	[ ]	*	9
(d)	72	=	9	*	[ ]

Figure 1: Equation formats for division problems in traditional format (a & b) conditions and recasted format (c & d). The missing "blank" element labelled as [ ] alternates between 3rd and 5th position in the equation.

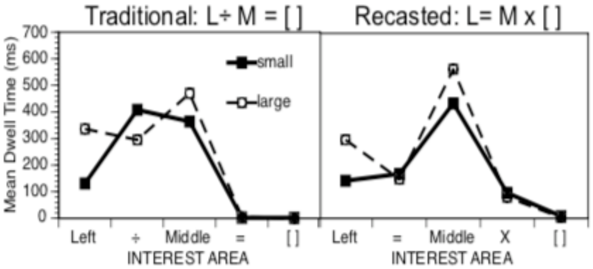

Figure 2: Mean gaze durations (total dwell time) for each symbol by format and problem size.

as compared to the other elements. In contrast, when participants were given a problem in the traditional format, their fixations were more evenly distributed across the problem elements (see Figure 2). Thus, analysis of fixation data showed that problem format influenced visual attention, which led the authors to speculate that mental recasting was taking place.

Present Work: Scanpath Analysis

The present work extends the Tan et al. (2016) results by analyzing the utility of scanpath data. For the scanpath analysis, we used an existing, freely available Python implementation of MultiMatch (https:github.com/adswa/ MultiMatch_gaze). Our primary aim was to investigate if there are differences in scanpaths between the traditional and recasted problem formats. If mental recasting is indeed occurring as previously hypothesized, we would expect the scanpaths to be affected by the problem format.

Pre-processing In preparation for data analysis, we first extracted the scanpaths from the eye tracker file, as follows. For each problem, we extracted the sequence of fixation co-ordinates and corresponding durations and stored them in a text file labelled with the participant and problem IDs (this step is a requirement of the MultiMatch tool). Thus, there was a single scanpath per problem. Scanpaths that had fewer than three fixations were not included (as required by the MultiMatch tool). Given that the vast majority of the *small* problems resulted in very few fixations, we only included the *large* problems in the present analysis. In summary, for each participant, we had multiple scanpaths for each problem format (on average about 30 scanpaths per format for each participant - the exact number varied slightly as even some of the large problems were solved in fewer than three fixations).

Scanpath Comparison Once the scanpaths were extracted, we followed the standard approach (Zhou et al., 2016) to compare scanpaths. This approach involves comparing scanpaths within each condition to each other and also comparing scanpaths between conditions to each other. The rationale is that if condition affects scanpath similarity, then the within-condition comparisons should produce higher similarity scores than the between-condition comparisons. In our study, we had two conditions corresponding to the two problem formats, traditional and recasted. If there indeed existed a difference in visual patterns on how these formats were processed, one would expect to see higher similarity for scanpaths corresponding to a given problem format than scanpaths corresponding to different formats. To obtain an overall measure of scanpath similarity, we wrote a Python script that called MultiMatch to calculate scanpath similarity scores using the approach outlined above[1]. Specifically, similarity scores were obtained for each participant for the following three types of comparisons:

(1) all unique pairwise comparisons of the traditional-format scanpaths (referred to as $\text{Intra}_{\text{traditional}}$ scores);
(2) all unique pairwise comparisons of the recasted-format scanpaths (referred to as $\text{Intra}_{\text{recasted}}$ scores);
(3) all unique pairwise comparisons of scanpaths *between* the two formats (referred to $\text{Inter}_{\text{traditional-recasted}}$ scores).

To briefly illustrate the process, the *intra* comparison involves *n choose k* scanpath comparisons, where *n* is the number of scanpaths in that condition and $k = 2$ given we are interested in pairwise comparisons (e.g., if there were only 4 scanpaths for a given format, this would produce $4!/4!(n-2)!$ similarity scores). The *inter* comparison involves *n1 x n2* scanpath comparisons, where *n1* and *n2* correspond to the number of scanpaths in each of the two collections. For every pair of scanpaths that were compared, MultiMatch produced five similarity scores, one for each MultiMatch feature (shape, direction, length, position, and duration). Scores were never averaged between features, due to the inherent difference in baseline similarity per feature and differences between methods used to measure the similarity. Once we had all the similarity scores for a given participant for each of the three types of comparisons ($\text{Intra}_{\text{traditional}}$, $\text{Intra}_{\text{recasted}}$, $\text{Inter}_{\text{traditional-recasted}}$), we obtained an average score for each feature and collection (i.e., there were five similarity scores for each participant for $\text{Intra}_{\text{traditional}}$, and likewise for

[1] Multimatch allows scanpaths to be grouped based on a number of parameters – this grouping parameter was set to false, as there currently is not sufficient understanding of when it is appropriate to use it.

Intra$_{recasted}$ and Inter$_{traditional-recasted}$). This approach is standard for scanpath comparison in a within-subject experimental design where participants are exposed to each study condition (Zhou et al., 2016).

Random Scanpaths To obtain a baseline for scanpath comparisons, we created a series of random scanpaths. In order to approximate the experimental setup, the following factors were taken into account when creating the random scanpaths: fixation coordinates, fixation duration, and scanpath length. Since each MultiMatch feature is calculated differently, random scanpaths demonstrated what similarity scores for each feature might resemble. With this purpose in mind, fixations were chosen across the entire visual space (rather than within the experimental AOIs). As per the average range found in the experimental scanpaths, we randomly varied the number of fixations for each random scanpath from 3 to 10. Fixation coordinates were assigned randomly for X(1 to 1600) and Y(1 to 1300), as per the coordinates of the screen on which experimental problems were presented. Fixation duration was assigned randomly from a range of 1(*ms*) to 3500(*ms*). This range of fixation durations corresponded to the range observed in experimental fixation durations. Thirty-three sets of data, each containing 60 randomly generated scanpaths (30 randomly labelled traditional format, 30 randomly labelled recasted format) were created to represent the number of participants in the study (*n*=33). These scanpaths were subject to the same comparison method as the participant scanpaths (for details, see *Scanpath Comparison* section).

Results

Unless otherwise stated, the results are based on data from the 33 participants in the original study. Results are only reported if they are significant ($p < 0.05$).

Does Problem Format Impact Scanpath Similarity?

As described above, data for the present analysis was obtained from division problems presented in the *traditional* and *recasted* formats. Did problem format influence patterns of visual attention? To answer this question, we analyzed the similarity of scanpaths, using the methodology outlined above. Recall that we had three collections of scanpath similarity scores, for each of the three comparison types described in the *Scanpath Comparison* section. If cognitive processes are indeed more similar within a condition, we would expect the *intra* comparison scores to be higher than the *inter* comparison scores.

Recall that MultiMatch produces a similarity score, which is a value between 0 and 1, for five different features per scanpath comparison. The descriptives for each feature are in Table 1, including the similarity scores for the random scanpath analysis, as well as the three types of comparisons (Intra$_{traditional,}$ Intra$_{recasted}$, Inter$_{traditional-recasted}$). The random scores serve as the baseline – their average similarity score ranges from .44 (*position*) all the way to .69 (*length*). Thus, MultiMatch produces fairly high similarity scores even when random scanpaths are compared. This was also reported by Dewhurst et al. (2012), who found that similarity for randomly generated scanpaths was high (0.64). This highlights the need for including a baseline benchmark in the analysis to ground the results, as we do here.

We followed up the descriptives with inferential statistics, using a one-way ANOVA with *comparison type* as the three level within-subject factor corresponding to the three types of similarity scores (Intra$_{traditional}$, Intra$_{recasted}$, Inter$_{traditional-recasted}$); the corresponding average similarity score was the dependent variable. Thus, in this analysis for each participant we had three similarity scores (i.e., Intra$_{traditional}$, Intra$_{recasted}$, Inter$_{traditional-recasted}$). Sphericity violations were corrected using the Greenhouse-Geisser adjustment. Significant effects were followed up with pairwise comparisons with Bonferroni correction.

There was a significant effect of *comparison type* on scanpath similarity for three MultiMatch features:

- *shape*, $F\ (1.5, 49.3) = 13.28, p < 0.01, \eta_p^2 = 0.29$
- *length*, $F\ (1.2, 38.6) = 11.29, p < 0.01, \eta_p^2 = 0.26$
- *position*, $F\ (1.6, 50.7) = 7.53, p < 0.01, \eta_p^2 = 0.19$

These results show that *comparison type* had a significant effect on scanpath similarity scores, but to see where that effect lies, follow up comparisons are needed. Here, the primary comparisons of interest are the two between condition comparisons (i.e., Intra$_{traditional}$ vs. Inter$_{traditional-recasted}$ and Intra$_{recasted}$ vs. Inter$_{traditional-recasted}$), because these tell us if

Table 1: Descriptives showing the similarity score for each MultiMatch feature, for the randomly generated scanpaths, followed by the two within-collection comparisons (traditional, recasted) and the between-collection comparison (traditional-recasted). Only one column is shown for the randomly generated scanpaths since as expected, the average similarity was virtually identical for all three comparisons.

	Similarity *random*		Intra Condition *traditional*		Intra Condition *recasted*		Inter Condition *traditional-recasted*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Shape	0.6598	0.0128	0.9448	0.0160	0.9387	0.0151	0.9391	0.0139
Direction	0.6391	0.0144	0.7687	0.1012	0.7707	0.0871	0.7599	0.0952
Length	0.6929	0.0176	0.9296	0.0184	0.9226	0.0149	0.9222	0.0145
Position	0.4389	0.0185	0.9004	0.0213	0.8965	0.0170	0.8918	0.0153
Duration	0.4983	0.0279	0.6256	0.0566	0.6370	0.0628	0.6287	0.0569

visual processing was affected by problem format. Specifically, if problem format affected scanpaths we would expect the *intra* scores to be higher for the target features than the *inter* scores. This is because the *intra* scores involve comparisons of scanpaths belonging to the same problem format (i.e., only traditional or only recasted), while the *inter* scores involve comparisons of scanpaths belonging to different formats (i.e., traditional vs. recasted). For the traditional format, this turned out to be the case for all three features, with the *intra* scores significantly higher than the *inter* scores (*shape*: $p < 0.01$; *length:* $p < .01$; *position*: $p < .01$). For the recasted format, this pattern held for one feature only, namely *position*, with the *intra* score significantly higher than *inter* score ($p < .01$).

The other pairwise comparison not yet reported analyzes if the average scanpath similarity within the traditional collection is different from the average scanpath similarity within the recasted collection. This can occur, for instance, if there is more variability in the scanpaths, leading to lower similarity scores. Prior work has indicated that lower similarity scores are obtained for more complex cognitive phenomena (Dewhurst et al., 2018). For the present data, the traditional format potentially requires the additional recasting step and so could be more complex than the multiplication problems that are already recast. If that were true, we could expect the similarity scores within the traditional format to on average be lower than the recasted format. On the other hand, the recasted format is less common and thus arguably more challenging – if that is the case, we could expect the similarity scores within this collection to on average be lower than within the traditional collection. The latter turned out to be the case for two of the three features under consideration, where the similarity scores for the $\text{Intra}_{\text{recasted}}$ were significantly lower than the $\text{Intra}_{\text{traditional}}$ scores (*shape*: $p < .01$; *length:* $p < .01$).

In summary, we found that scanpath similarity was affected by problem format for three MultiMatch features, namely *shape*, *length*, and *position*.

Follow-up Analyses

We conducted several follow up analyses. First, we verified that there were no significant differences between the three comparison groups for the random scanpaths. As expected, none of the main effects between condition comparisons were significant ($p > .14$), providing credibility for the MultiMatch approach. Second, we verified that the presentation format did not influence results.

The traditional and recasted problems in the original study varied the position of the missing element in the equation, i.e., "blank". which was the placeholder for the solution (see [ ] in Figure 1). To identify if the position of the blank influenced outcomes, we labelled the scanpaths with the location of this element and re-ran the similarity MultiMatch analysis. We then added a second two-level factor to the ANOVA (position_1, position_2 corresponding to the position of the blank) and re-ran the inferential statistics. The main effect of blank position was not significant for any of the five features ($p > .139$), while the effect of condition remained significant and held the same pattern as above (effect of *shape, length,* and *position* features resulted in significant main effects, $p < 0.05$).

Third, we checked the effect of latency. In the original experiment, participants were asked to generate their solutions as "quickly and accurately" as possible. If participants answered quickly, it was more likely they were retrieving the solution from memory directly as opposed to recasting it (in theory, retrieval is possible even in recasted format). If participants solved problems primarily using retrieval, then there would be no differences in scanpath similarity scores between traditional and recasted formats, because retrieval does not require scanning and/or shifting visual attention between problem elements,

To check the effect of solution latency, we first identified the median solution time. Based on a median split, we then labelled scanpaths' as slow (longer than 1 second) or fast (equal to or slower than 1 second). Because not all participants had both types of scanpaths, we conducted two separate one way ANOVAs with comparison type as the factor, one ANOVA for "fast" scanpaths and one for the "slow" scanpaths. We then aggregated the data as for the primary analysis, by obtaining for each similarity collection ($\text{Intra}_{\text{traditional}}$, $\text{Intra}_{\text{recasted}}$, $\text{Inter}_{\text{traditional-recasted}}$) the mean score for slow scanpaths and the mean score for fast scanpaths.

As we anticipated, none of the three main effects were significant for the fast scanpath analysis. For the slow scanpath analysis, we were left with 25 participants (as some participants only had fast responses). We found the similar pattern of results as for the primary analysis; with significant main effects for *shape*, *length*, and *position*, the being caveat that some of the follow up comparisons were no longer significant.

Discussion

The goal of the present study was to analyze whether scanpaths are affected by problem format. Two types of formats were included, namely a traditional division format (72 ÷ 9 = [ ]) and a recasted multiplication format (72 = 9 × [ ]). Prior work suggests that when participants are given a problem in the traditional format, they first mentally recast the problem into the multiplication format and then produce the answer (this is particularly the case if the division format includes large values). In contrast, for problems in the recasted format, participants directly solve the problem without mental recasting. Thus, the proposal is that problem format affects the strategy used to solve the problem. Evidence for this conjecture in prior work came from (1) latency differences, namely that problems in the traditional format took longer to solve (because they require the recasting step) and (2) fixation analysis of eye data, which showed that the distribution of dwell time over the corresponding problem elements was different between the two formats. This provides some indication that, as hypothesized, there are strategy differences between the two formats. We provide further evidence that problem format may affect strategies, by showing that format elicited significantly different scanpaths for three MultiMatch features (*shape*, *length*, and *position*). *Shape* and *position* features capture the similarity related to the order of fixation locations within a scanpath. The higher similarity scores for the traditional problem-format scanpaths suggest increased consistency of scanning patterns for traditionally-formatted division problems. Traditional division problems follow a standard presentation of the equation (see Figure 1), which may encourage a left to right reading pattern. In contrast, the recasted multiplication format presents the solution on the left-hand side of the equation. This atypical presentation could impact the default reading strategy and result in increased variation of viewing patterns, reducing similarity scores (this is indeed what we found for the shape feature). Higher similarity for the *length* feature for the traditional format scanpaths also suggests a standard reading pattern is present, since saccade distances, if the eye is moving from one symbol to the next in the equation, should remain relatively similar in length.

The *duration* and *direction* features were not informative in the present analysis, perhaps because there was not a great deal of variability in terms of these variables in the scanpaths due to the experiment's relatively simple stimuli. These two features might be of more value in studies that use more complex visual stimuli. For instance, prior work has indicated differences between novices and experts for fixation duration and perhaps these differences would also show up in the duration scanpath feature (Reingold et al., 2001; Balslev et al., 2012; Susac et al., 2014). In general, as noted in (Jarodzka et al., 2010; Dewhurst et al., 2012), there currently do not exist formal guidelines about the usage and interpretation of MultiMatch features, and so work is needed in this area.

Our research makes both a theoretical and practical contribution. On the theoretical side, we provide additional evidence that problem format produces different patterns of visual attention. Specifically, here we extend prior work using fixation data (Tan et al., 2016) with scanpath analysis. On the practical side, our results are aligned with previous studies reporting differences in scanpaths between experimental conditions (Zhou et al., 2016). However, to the best of our knowledge, our work is the first to target basic arithmetic problem solving in the context of scanpath analysis. Our results provide initial evidence towards the utility of scanpath analysis in math cognition domains but more work is needed.

We do acknowledge that scanpath analysis is not without limitations. One limitation, highlighted by our overview of related work, is that scanpath analysis can tell us whether differences in scanpaths exist between conditions but not much beyond that, rendering the underlying cognitive processes as a black box. A step in addressing this limitation relates to identifying representational scanpaths to see what patterns of visual attention look like, as was done in by Zhou et al., (2016). The challenge with this approach is that the scanpaths quickly become unintelligible as the number of fixations in them increases. Thus, visualization tools are needed to shed light on the structure of scanpaths.

Another challenge and associated limitation pertains to the methodology underlying scanpath analysis. This analysis is inherently relational in nature, meaning that conditional differences can only be identified by comparing two or more groups of similarity scores. Doing so becomes more complicated if a condition involves multiple trials, for instance to control for variability, as was the case in the decision-making studies reported by Zhou et al. (2016) and the math cognition task used here. Including multiple trials requires a great deal of scanpath comparisons to generate the corresponding similarity scores.

In conclusion, while we are cautiously optimistic about the utility of scanpath analysis, given that this approach is relatively new, guidelines on choosing appropriate methodologies and tools for various cognitive domains are needed.

Acknowledgements

This research was supported by an NSERC Discovery Grant.

References

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior research methods*, *50*(1), 362-379.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior research methods*, *42*(3), 692-700.

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior research methods*, *44*(4), 1079-1100.

Foerster, R. M., & Schneider, W. X. (2014). Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*, *8*(2), 6-6.

French, R., Glady, Y. & Thibaut, J. (2017). An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making. *Behavior research methods*, *49*(4), 1291-1302.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*, 1–58.

Holmqvist, K., Andrà, C., Lindström, P., Arzarello, F., Ferrara, F., Robutti, O., & Sabena, C. (2011). A method for quantifying focused versus overview behavior in AOI sequences. *Behavior research methods*, *43*(4), 987-998.

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. Learning and Instruction, 20(2), 146-154.

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instructional Science*, *40*(5), 813-827.

Lai, M. L., Tsai, M. et al. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational research review*, *10*, 90-115.

LeFevre, J. A., & Morris, J. (1999). More on the relation between division and multiplication in simple arithmetic: Evidence for mediation of division solutions via multiplication. *Memory and Cognition*, 27, 803-812.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics – Doklady, 10, 707–710.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. Spatial Vision, 9, 363–386.

Nivala, M., Hauser, F., Mottok, J., & Gruber, H. (2016, April). Developing visual expertise in software engineering: An eye tracking study. In *2016 IEEE Global Engineering Education Conference* (pp. 613-620). IEEE.

Reingold, E., Charness, N., Pomplun, M., & Stampe, D. (2001).Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*(1), 48-55.

Sharafi, Z., Soh, Z., & Guéhéneuc, Y. G. (2015). A systematic literature review on the usage of eye-tracking in software engineering. Information and Software Technology, 67, 79-107.

Susac, A. N., Bubic, A., Kaponja, J., Planinic, M., & Palmovic, M. (2014). Eye movements reveal students' strategies in simple equation solving. *Int. Journal of Science and Mathematics Education*, *12*(3), 555-577.

Tan, S., Muldner, K., & LeFevre, J. (2016). Solution of division by access to multiplication: evidence from eye tracking. *In Proceedings of the Cognitive Science Society Conference*, 656-661.

Von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of memory and language*, *94*, 119-133.

Wilson, K. A., Heinselman, P. L., & Kang, Z. (2016). Exploring applications of eye tracking in operational meteorology research. *Bulletin of the American Meteorological Society*, *97*(11), 2019-2025.

Zhou, L., Zhang, Y. et al. (2016). A scanpath analysis of the risky decision-making process. *Journal of Behavioral Decision Making*, *29*(2-3), 169-182.