How task demands influence scanpath similarity in a sequential number-search task

Richard Dewhurst[a,e,*], Tom Foulsham[b], Halszka Jarodzka[c], Roger Johansson[d], Kenneth Holmqvist[f,g,h], Marcus Nyström[e]

[a] *Århus University, Interacting Minds Centre, Århus, Denmark*
[b] *University of Essex, Department of Psychology, Colchester, U.K*
[c] *Open University of the Netherlands, Welton Institute for Learning, Teaching and Technology, Heerlen, Netherlands*
[d] *Lund University, Department of Psychology, Lund, Sweden*
[e] *Lund University Humanities Lab, Lund, Sweden*
[f] *Department of Psychology, Regensburg University; Germany*
[g] *Department of Computer Science, Bloemfontain University, South Africa*
[h] *UPSET, Northwest University, South Africa*

Abstract

More and more researchers are considering the omnibus eye movement sequence—the scanpath—in their studies of visual and cognitive processing (PloS One 6 (2011) e18262; Journal of Memory and Language 65 (2011) 109-127; Journal of Vision 11 (2011) 1-11; ETRA Proceedings (2012) 193-196). However, it remains unclear how recent methods for comparing scanpaths perform in experiments producing variable scanpaths, and whether these methods supplement more traditional analyses of individual oculomotor statistics. We address this problem for MultiMatch (ETRA Proceedings (2010) 211-218; Behavior research methods 44 (2012) 1079-1100), evaluating its performance with a visual search-like task in which participants must fixate a series of target numbers in a prescribed order. This task should produce predictable sequences of fixations and thus provide a testing ground for scanpath measures. Task difficulty was manipulated by making the targets more or less visible through changes in font and the presence of distractors or visual noise. These changes in task demands led to slower search and more fixations. Importantly, they also resulted in a reduction in the between-subjects scanpath similarity, demonstrating that participants' gaze patterns became more heterogenous in terms of saccade length and angle, and fixation position. This implies a divergent strategy or random component to eye-movement behaviour which increases as the task becomes more difficult. Interestingly, the duration of fixations along aligned vectors showed the opposite pattern, becoming more similar between observers in 2 of the 3 difficulty manipulations. This provides important information for vision scientists who may wish to use scanpath metrics to quantify variations in gaze across a spectrum of perceptual and cognitive tasks.

Keywords: Scanpaths, Eye Movements, MultiMatch, Visual Search

*Corresponding author

1. Introduction

Where we direct our eyes, when and for how long often serve as direct behavioral correlates of attentional selection. Because of this link between where a person is looking and how they are currently processing information, the study of eye fixations and saccades during simple and complex tasks has become commonplace (e.g. Rayner, 2009). However, individual oculomotor statistics analysed separately do not always provide the fullest picture of the eye movement behaviour elicited by observers; hence the proliferation of scanpath visualizations in papers to support arguments based on isolated oculomotor statistics alone (Figure 1). A combined representation can be very useful and so, to accommodate the trend of referring to scanpath visualisations in the literature there has been something of an explosion of scanpath comparison metrics in recent years (Anderson et al. (2015); and other recent developments, e.g. Kübler et al. (2016); Wilson et al. (2018)). These help quantify our intuitive sense of similarity/dissimilarity present in figures representing eye movement sequences. Nevertheless, scanpath comparison methods are often described in methods papers (e.g. Cristino et al., 2010; Dewhurst et al., 2012; Foerster & Schneider, 2013), or used with high level cognitive tasks where it can be hard to pinpoint the type of similarity identified and interpret its meaning (Goldberg & Kotval, 1999; Foulsham et al., 2012). We sought here to evaluate our MultiMatch method with a simple perceptual task, more alike those often employed in vision research. Given MultiMatch's growing popularity and potential (French et al. (2016) for instance, recently found our method to be the "..most efficient one for examining scanpaths during analogy making", p. 9), we hope this will provide useful information for basic vision scientists as well as those using scanpath comparison techniques in more applied domains, where comprehensive evaluation is lacking.

[INSERT FIGURE 1 HERE]

1.1. Single oculomotor statistics: The components of a scanpath

Researchers commonly report effects on single eye movement events, which may or may not be tied to areas of interest (AOIs) in the stimuli. For example, it is common practice to see a table of standard oculomotor statistics in journal papers consisting of number of fixations, fixation or dwell duration, and saccadic amplitude or length etc. (e.g. Dewhurst & Crundall, 2008; Rayner et al., 2008; Foulsham & Kingstone, 2010). Such statistics are relevant for a number of reasons. First, measuring the number of fixations, gives an indication of the number of shifts of attention necessary to complete the task. As such, this measure often correlates very highly with the amount of time spent on a task (or the reaction time where this duration is determined by a response). For example, slower, more difficult search trials will result in longer reaction times and, often a greater number of fixations. This pattern can be used to argue that search gets slower because of changes in attentional selection of distractors and targets, rather than, for example, a change in response criteria.

Second, measuring the time spent processing individual items by calculating fixation or dwell time (where dwell is the sum duration of consecutive gaze on an item) is assumed to largely reflect the difficulty of processing stimuli at the fixated location. The word frequency effect (Rayner & Raney, 1996), the effect of informational load (Gould, 1973), and in usability studies, the difficulty of extracting information from a display (Goldberg & Kotval, 1999), all support the general finding of longer fixation or dwell times as a function of greater cognitive demands. As a result, these measures of processing time are expected to accord to how easily an item can be apprehended by

the visual-cognitive system. Within a fixed time limit, the number of fixations and their duration are inversely proportional.

The above ocolomotor statistics are bound together by the third eye movement measure: saccadic amplitude, or length. Because larger saccades target items further from the high-resolution fovea, they reflect an ability to detect peripheral features more easily. In contrast, a hard task would be expected to lead to smaller saccades because more distractors have to be inspected, or because of a reduced perceptual span; with more difficult tasks the area around fixation from which we can extract information shrinks (Reingold et al., 2001; Pomplun et al., 2001).

One of the challenges of eye movement research is that there are a large number of derived measures beyond these three basic statistics. In simple tasks, there may be very straightforward measures which represent the behaviour of interest (e.g., saccade gain in a target step task). In less constrained situations, there are potentially many "researcher degrees of freedom" regarding which measure to use, and the chosen statistic should be determined by theory and the predicted behaviour of interest. Scanpath analysis potentially presents an attractive alternative because it can summarize the general pattern of viewing over time. In the present study we consider whether MultiMatch can summarize different patterns of behaviour in response to changes in task difficulty. Since MultiMatch provides a number of measures quantifying scanpath similarity, it can also provide flexibility for addressing specific questions.

Relying on individual summary statistics in isolation may also miss out on important relationships between different aspects of eye movements. For example, fixation duration and saccade amplitude may covary during scene viewing (Unema et al., 2005). As viewing progresses through an image, fixations tend to become longer in duration and saccades become smaller in amplitude. It has been argued that this relationship tells us something specific about the attentional processing happening at each point in time (with focal processing increasing as viewing goes on). Similarly, Wilson et al. (2018) show that where no significant differences are found in fixation measures of experienced weather forecasters viewing a radar display, there are reliable differences in their scanpath similarity scores identified with MultiMatch. However, we are not suggesting here that our method, or indeed scanpath measures generally, are a *substitute* for single oculomotor statistics or what can be derived from them; rather, we wish to point that MultiMatch is complimentary to more traditional eye movement measures. It cannot it its own right say anything about underlying changes in eye movement behaviour, only changes in the similarity of eye movement sequences. But it does have the novel ability to reveal hidden attributes within a chain of temporally connected fixations, and inferences can be made about mechanisms when used in combination with other eye movement statistics.

When considered en masse the scanpath can reflect changes in perceptual or cognitive demands which may be hidden if individual measures are taken one at a time. In the experiment described here, by objectively controlling the source of task difficulty perceptually while keeping the cognitive task constant, we shed light upon the common properties shared between observers' scanpaths for the kind of task which has traditionally relied upon individual eye movement statistics, analysed separately. Operationalizing the task in this way allows us to identify which features of the whole scanpath are driven by task difficulty and aspects of the display.

1.2. The present study

In this experiment we investigate the between-subject similarity of gaze behaviour across several conditions. Starting with an experimental task where we expect people to perform in a highly similar fashion to each other (looking at five items in a prescribed order), we ask how we can represent this

similarity using individual oculomotor measures (number of fixations, fixation duration and saccadic amplitude) as well as the omnibus eye movement scanpath. We use our scanpath comparison measure MultiMatch (Jarodzka et al., 2010; Dewhurst et al., 2012) which quantifies our intuitive sense of similarity very well, and provides a number of dimensions of similarity rather than just one overall score. We then increase the difficulty of the task, in order to test the prediction that participants will become more idiosyncratic as there becomes more room for errors. Our main aim was to understand the similarity of scanpaths, and in particular the behaviour of vector-based similarity metrics like MultiMatch, in a constrained visual task. By investigating general oculomotor statistics and behavioural performance alongside the whole scanpath representation, we provide insight of eye guidance and scanpath similarity in this task. This will provide insights into the aspects of gaze behaviour which are common to participants under particular conditions, and those that are more variable. It will also provide a test for MultiMatch, and attempts to quantify scanpaths more generally.

The basic experimental paradigm presented participants with the numbers 1–5 in random locations on the screen. Their task was to saccade to, and fixate each number in numerical order, thus producing eye movement sequences—scanpaths. Three manipulations of this basic task were implemented in order to investigate whether scanpath similarity between participants is critically related to different levels of crowding and conspicuity. First, the numbers could be presented in different font sizes. Second, they could be shown along with distractors; that is, varying set size. Third, the numbers themselves remained unchanged, but the background noise level was systematically altered. Figure 2 illustrates examples of each case.

These three conditions each contained five levels of difficulty, where the font size became increasingly smaller, the number of distractors increased, or the background noise intensified, in five steps, making the exercise of locating the numbers in the right order more challenging. When each number is highly visible, we would expect both the location and order of fixations to be similar between observers. Under these conditions, participants should find it easy to locate the next target (e.g., in peripheral or parafoveal vision), reducing the need to look at other locations or back to previous targets, and producing a scanpath which approaches the ideal sequence from 1 to 5. As the visibility of the target numbers decreases, we expected participants to produce more divergent scanpaths, both from this ideal sequence and from each other. While the task requires that the same locations be inspected, participants should find it more demanding to fixate these targets in the correct order, leading to a different sequence of gaze behaviour. Our analysis aims to uncover which aspects of the scanpath change under such conditions (for example, the duration of fixations, their precise spatial position, or the angle of the saccades involved). As the task becomes more difficult, we expect more strategic or random influences on individual participants, and thus we predict decreasing between-subject scanpath similarity.

2. Method

2.1. Participants

Twenty participants (9 female, 26.9±5.3 years of age) participated in the experiment. All participants had normal or corrected-to-normal vision. They were mainly recruited from the general student population at Lund University, and were re-imbursed with a lottery scratch card for their time.

2.2. Software & Apparatus

Stimuli were displayed on a Samsung Syncmaster 931c TFT LCD 19 inch (380×300 mm) screen running at 60 Hz, with a resolution of 1280×1024 pixels. The experiment was run in Matlab R2009b using the Psychophysics toolbox Brainard (1997). Binocular eye-movements were recorded at 500 Hz with an SMI HiSpeed tower system, and iView X 2.5.

2.3. Stimuli & Design

150 stimulus images were generated where the numbers `1,2,3,4,5` were presented either by themselves, together with additional numbers irrelevant for the task, or embedded in noise. To systematically vary the difficulty of the task—looking at the numbers 1-5 in increasing order—each of the three presentation alternatives was manipulated with respect to *font size*, *set size* (number of distractors), and the level of *background noise*. Details about the manipulations are as follows (all these values were set by pilot testing):

- **Font size** Five different font sizes were used; $\exp(x)$, where $x \in \{2.50, 2.85, 3.20, 3.55, 3.90\}$. This equates to a size range of between 12pt–49pt on screen.
- **Set size** The number of distractors n varied between 1 and 5, and added the additional numbers $5 + \{1, 2, \ldots, n\}$ to the stimulus display. Where the numbers 1–6 were presented this therefore relates to the minimum set size of 6 (i.e. 1 distractor), up to a maximum set size of 10 (i.e. 5 distractors) where the numbers 1–10 were displayed.
- **Noise level** Noise levels λ were chosen from the set $\{12, 37, 63, 88, 114\}$. If

$$\begin{aligned} I_N(x,y) &= 128 + I_z(x,y) \\ I_z &\sim \mathcal{N}(0,\lambda) \\ x &= 1,2,\ldots,M \\ y &= 1,2,\ldots,N \end{aligned} \tag{1}$$

  denotes a noise image and $I_D(x,y)$ represents a midgray image with black numbers superimposed, then

$$I_S(x,y) = \alpha I_D(x,y) + (1-\alpha) I_N(x,y) \tag{2}$$

  defines the final stimulus image with $M = 1280, N = 1024$ and $\alpha = 0.5$. The noise manipulation therefore did not alter the number or size of the numbers themselves, but rather how distinghuishable they were from the background, in five exponential steps.

Ten images were generated for each manipulation type and level, yielding 150 trials (3 conditions $\times$ 5 difficulty levels $\times$ 10 images).

Numbers were shown in Courier New font, and were positioned in the centre of a randomly chosen section of an invisible 5×5 grid dividing stimulus-space. To prevent participants from perceiving that numbers were presented overlaid on a grid, an additional offset from the interval $d \in [0, 2]$ degrees was added to the position of each number away from the centre of the grid section it occupied. Only one number could be assigned to each of the 25 sections of the grid.

Figure 2 illustrates the three conditions, with examples of stimulus images for trials at different difficulty levels.

[INSERT FIGURE 2 HERE]

2.4. Procedure

Informed consent was obtained from each participant when they arrived to the Humanities Lab, and the experiment was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Instructions were provided on the monitor explaining the task, and the experimenter was present throughout to clarify any uncertainties. Once having agreed to take part and confirming they understood what was required, a 13–point calibration procedure followed on the HiSpeed system. Directly following this, the four points oblique to the centre were re-presented for validation. Validation accuracy across all participants was 0.90 ± 0.60 degrees (x offset), and 0.60 ± 1.00 degrees (y offset) (M±SD). Viewing position was stabilized at 67cm throughout using SMI's built in chin-rest.

After calibration (and validation), participants were initially given a short practice block consisting of the presentation of 10 sequences, randomly drawn from the block which was about to follow. Then participants commenced the experiment proper.

Each trial began with a central fixation cross for 2000ms. Next, when presented with the numbers, participants were asked to look at the centre of each (1–5) in turn in increasing numerical order, and press the left mouse button when reaching the number 5. The mouse click ended each trial and triggered the presentation of a new trial. Numbers larger than 5 were to be ignored (in the set size case). The stimulus types of the three conditions (font size, set size, noise level) formed one block each, and images were selected randomly, without replacement, until all 50 for the block in hand had been shown. Block order was also randomized between participants.

Participants were asked to perform the task as quickly but as accurately as possible. A break in between blocks was provided if requested, where re-calibration was carried out if necessary. Depending on participant variation and calibration time, one testing session lasted approximately 40mins. As all participants viewed every unique number arrangement for each condition, this allowed us to assess scanpath similarity between-subjects as a function of perceptual task difficulty (i.e. the similarity between two participants' scanpaths when viewing the same stimulus array, easy or more difficult).

2.5. Data pre-processing

2.5.1. Eye movements

Oculomotor events were estimated from raw data samples using Nyström & Holmqvist's 2010 adaptive algorithm for fixation, saccade and glissade detection (with the standard settings stated in their paper). The first fixation in each trial was discounted from subsequent analysis since participants always began looking at the initial central fixation cross.

For calculation of task performance and exclusion criteria, first AOIs were placed encompassing each number within a square perimeter as tightly as possible, to which a margin of 2° was added to account for minor spatial offsets in the data, and/or failure of the participant to look directly at the number. If for some reason a fixation would be located in two AOIs simultaneously (which is highly unlikely given the distribution of the numbers in space), the AOI with the smallest fixation-AOI centre distance was chosen.

2.5.2. Scanpaths

The MultiMatch algorithm, full details of which are explained elsewhere (Jarodzka et al., 2010; Dewhurst et al., 2012), revolves around the principles of first simplifying scanpaths into virtualised sequences of saccadic vectors, then aligning one scanpath with another on the basis of their shape. From here calculating similarity is a simple matter of subtracting the dimensions between aligned

vector pairs for the whole sequence, and taking an average. The five resulting dimensions of similarity are: Shape (vector difference between aligned saccade pairs), Length (difference in amplitude between the endpoints of saccade vectors), Direction (angular difference in heading), Position (Euclidean difference in x-y locations of aligned fixations) and Duration (difference in the length of time elapsed between fixation pairs). Figure 3 depicts this method.

One participant was excluded due to having very low behavioural accuracy compared to the others (longest common subsequence, described below, < 1 on average across all conditions and trials).

[INSERT FIGURE 3 HERE]

3. Results

How do participants fare at this task? What strategy do they use? The results are broken down into three main sections to tackle such questions. First, we present behavioural task performance to address whether the difficulty manipulation was effective—we expected longer search times and decreasing accuracy as the task becomes harder. The second section of the results concentrates on the individual eye-movement components of a scanpath—number of fixations, fixation duration, and saccadic amplitude. Third, we present the scanpath comparison results, exploring the similarity scores returned by MultiMatch, and whether these are lower when the task is harder, as we predicted. In this way we can decompose the scanpath, and associated eye movement behaviour, so as to better understand scanpath similarity and the kinds of results produced by MultiMatch with eye movement data typical of experiments on visual perception and cognition.

3.1. Task performance

3.1.1. Search time

The time taken to complete a trial was longer with smaller font sizes, with larger set sizes (more distractor numbers), and with higher noise levels (see figure 4). These results were confirmed with a repeated-measures ANOVA with two factors (condition and difficulty level). This analysis revealed a significant main effect of condition ($F_{2,36} = 10.02, MSE = 1.79, p < .001$), indicating that font size ($\bar{X} = 3960ms$) was completed more quickly than the set size ($\bar{X} = 4770ms$) or noise level ($\bar{X} = 4630ms$) conditions. This finding was supported by bonferroni-corrected post-hoc comparisons (p's <0.01). It is most likely that this result is owing to the fact that, overall, the numbers are larger in the font size condition. An interaction was also observed ($F_{8,144} = 27.89, MSE = .133, p < .001$), showing that while search time increased as a function of increased set size or noise level, the pattern was reversed for the font size condition. This is to be expected, since as the numbers become larger they are *easier* to locate. In terms of task demands therefore, the difficulty of the oculomotor exercise had a linear effect on search times; harder trials being associated with extended search, as predicted. This was verified with interaction contrasts, where each level of difficulty is always significantly different from the previous levels (all ps $<.001$) [1]

[1] As any interactions translate as a direct effect of task difficulty in this way (i.e. due to its inversion for font size), we henceforth just refer to a 'linear effect of difficulty'.

3.1.2. Accuracy

Performing the task accurately required fixating the five numbers in order. To quantify this, we used the Longest Common Subsequence (LCS, Hirschberg, 1977). This was defined as the length of the longest sub-sequence common to the 'ideal' and observed scanpaths. According to this definition, each empirically-observed sequence of fixated targets was evaluated against a hypothetical 'ideal viewer' (i.e., taking the correct path directly, without errors; [1, 2, 3, 4, 5]). For instance, the observed scanpath [1, 2, 1, 3, 2, 5] would give a LCS of 4, since the sub-sequence 1235 is common to both scanpaths. A high degree of commonality between a participant's observed scanpath and that of the ideal viewer indicates the number sequence is being followed correctly, and will return a high LCS value, approaching the maximum of 5. Our use of the LCS is similar to the sequence comparison methods used elsewhere in eye movement research (i.e., the string edit distance; Foulsham & Underwood (2008)). Importantly, although this measure captures the overall instructions for the task, it does not consider global differences, detours or repetitions from the ideal scanpath. An average LCS close to five reflects participants performing well. Lower scores indicate errors, such as mistakenly fixating the wrong number, forgetting where one is up to in the sequence or skipping a number. If the LCS decreases with difficulty then it would demonstrate that participants find the task more visually challenging. Because it is possible for two participants to perform equally well (high LCS) but have very different scanpaths (e.g., because one has many deviations), the LCS should be considered in conjunction with the search time data (above), and the other oculomotor statistics in the next subsection.

Figure 5 shows LCS scores for all participants, broken down over condition and difficulty level. The LCS scores changed with difficulty, particularly in the font size and set size conditions, but there was a main effect of condition ($F_{2,36} = 65.29, MSE = .235, p < .001$), indicating superior performance with set size ($\bar{X} = 3.69$), compared to font size ($\bar{X} = 2.96$) or noise level ($\bar{X} = 3.02$). A linear effect of difficulty was also observed (interaction: $F_{8,144} = 60.31, MSE = .058, p < .001$), but surprisingly this was in the opposite direction from predictions: Participants became less accurate with larger fonts and more accurate with larger set-sizes.

[INSERT FIGURES 4 + 5 HERE]

3.1.3. Discussion

We hypothesized that if the relationship between our task difficulty manipulation and behavioural performance was straightforward we should see a corresponding decrease in accuracy as difficulty increases, coupled with an increase in search times. This was not the case; set size particularly producing significantly *improved* performance at the harder end of the difficulty scale. It can be seen from the search time results that this result may in part owe to a speed accuracy trade-off, present in the font size condition and more pronounced (in terms of accuracy) in the set size condition. What is it about these stimulus types that encourages this type of search behaviour? It is likely that the larger (easier) numbers can be seen via peripheral vision, so targeting a fast saccade to a known number location is not costly to execute, even if the exact identity of the number is not always known in advance. Indeed, it was in this condition where the steepest slope for search time as a function of difficulty was observed. Set size, conversely, induces a steeper slope for accuracy, possibly due to participants modifying their default search strategy to account for the larger number of potentially correct targets which cannot be discerned outside of foveal vision. One could hypothesize that they slow at larger set sizes to allow for more precise saccade targeting, which is more efficient in this case. We turn now to the oculomotor data to shed further light on these possibilities evident in participants eye movements.

3.2. General oculomotor behaviour

How do participants' eye movements compare to their behavioural performance? To tackle this question we analysed the number of fixations per trial, fixation duration, and saccade amplitude. As discussed in the introduction, these eye movement variables comprise the scanpath and are all shown to change in different ways as a function of the difficulty of the task being carried out. Importantly, contrasting outcomes between these conditions in terms of number and duration of fixations, as well as saccadic amplitudes, would produce quite different effects on each of MultiMatch's dimensions (*shape, length, direction, position and duration*).

3.2.1. Number of fixations

The number of fixations per trial showed the same pattern as search time (see figure 6). There was again a main effect of condition ($F_{2,36} = 65.29, MSE = .235, p < .001$), with the set size condition ($\bar{X} = 12.26$) giving the highest number of fixations, followed by noise ($\bar{X} = 10.75$), then font size ($\bar{X} = 9.46$). Bonferroni-corrected post-hoc tests confirmed this (all *ps* <.005). There was also a significant linear effect of difficulty paralleling the search time results (interaction: $F_{8,144} = 34.88, MSE = 1.39, p < .001$), showing that fixations are more numerous at harder difficulty levels. This finding was supported by interaction contrasts, where each level was significantly different from the previous levels (all *p*'s <.005). This mirrors the previous analyses.

Note that these data compliment the LCS accuracy results well. The average number of fixations per trial, overall, was 10.8 ± 3.35. This means that participants made around 5 extra fixations from the ideal 5 if their accuracy was perfect; but it is likely a few more than this occurred since accuracy was not at ceiling. In any case we can be confident that the number of fixations leading to correct task performance was not excessive, and that the scanpaths contain some variability which we can quantify.

[INSERT FIGURE 6 HERE]

3.2.2. Fixation duration

Unlike the previous measures, task difficulty had a much less systematic effect on fixation duration. The ANOVA only revealed a marginally significant linear effect of difficulty for the fixation duration analysis (interaction: $F_{8,144} = 2.13, MSE = 0.00, p = .037$)(see figure 7). If anything, fixations trend to be slightly shorter with increasing difficulty in the font and set size conditions. But as the assumption of sphericity was violated and the Greenhouse-Geisser corrected p value equals .084, the validity of this finding is questionable. Thus there was little evidence for a change in fixation duration across conditions.

[INSERT FIGURE 7 HERE]

3.2.3. Saccadic amplitude

As we have seen in the introduction, saccade amplitude is also seen to vary with task difficulty, reflecting the ability to detect task relevant items in peripheral vision. Saccades are also the composite feature of scanpaths, combining fixations into the whole.

Generally, saccadic amplitudes were longer when the task was easier; that is, with larger numbers, fewer distractors, and less background noise (see figure 8). The same analysis was once again carried out. This ANOVA revealed a significant main effect of condition ($F_{2,36} = 12.32, MSE = 1.182, p < .001$), indicating that saccade amplitudes were larger in the font size condition ($\bar{X} = 8.6°$)

compared to the set size condition ($\bar{X} = 7.8°$). This finding was supported by bonferroni-corrected post-hoc comparisons, where these conditions were reliably different ($p < .001$). Noise level fell in-between ($\bar{X} = 8.2°$). This nicely demonstrates that, irrespective of difficulty, eye movements select the font size number stimuli (which are on average larger) more directly, with fewer small shifts of attention in between. When there are distractors however (and the numbers are on average smaller), one cannot disambiguate targets from distractors in peripheral vision as easily, and saccades become less efficient, not heading straight for the right numbers in turn. This fits well with known effects of crowding (e.g. Vlaskamp & Hooge, 2006). One should note here that this is a good example of where one metric—in this case saccadic amplitude—does not by itself provide an entirely conclusive case of the interpretation drawn. Yes, it is likely that search has become less efficient in the presence of distractors, but in order to be sure of this and try to understand exactly why, we need another complimentary tool. In this paper we chose scanpath comparison with our own algorithm, but there are many other measures available (see Holmqvist et al., 2011), and the choice should be guided by what best answers your research question rather than what is most readily available.

There was also a significant overall (i.e irrespective of the interaction term brought about by the inversion of difficulty for font size; see p. 7) main effect of difficulty level ($F_{4,72} = 18.80, MSE = .203, p < .001$). Difficulty levels 4 and 5 gave rise to shorter saccades ($\bar{X} = 8.0°$, and $\bar{X} = 7.8°$ respectively) compared to the previous levels (both ps <0.001). This shows that, overall, smaller saccades are associated with increasing difficulty and vice versa, in line with previous findings.

There was also a linear effect of difficulty revealed in the interaction term ($F_{3.95,71.07} = 10.22, MSE = .704, p < .001$). Degrees of freedom were adjusted here according to Greenhouse-Geisser due to a violation of the assumption of sphericity. Interaction contrasts showed both levels 3 and 4 were significantly different from the previous levels (at p <.001), as was difficulty level 5 (at p <.05), supporting the approximately linear interpretation.

[INSERT FIGURE 8 HERE]

3.2.4. Saccadic targeting

To get qualitative insight into how task difficulty influences saccade targeting, "heat maps" are used. The heat maps were generated by superimposing two-dimensional Gaussian functions, each centered at participants' fixation locations. Figure 9 shows six randomly selected trials for the largest (top row) and the smallest (bottom row) font sizes. The black circles indicate where the numbers were located. More distributed saccade landing points for the largest font size would lead to softer peaks in the heat maps, which seems not to be the case.

[INSERT FIGURE 9 HERE]

In Figure 10, heat maps from six randomly selected trials for the largest ($n = 10$, 5 distractors, top row) and smallest ($n = 6$, 1 distractor, bottom row) set sizes are shown. Distractors are marked with red plus signs ('+'). It is directly evident from the figure that while the targets are the most frequent fixation targets, the distractors are also fixated occasionally. Fixations to spaces in-between items seem sparse, unless the items are in close proximity. The figure does not seem to support the hypotheses that more distractors lead to more precise saccade targeting (sharper peaks in the heatmap).

[INSERT FIGURE 10 HERE]

3.2.5. Discussion

With regard to the font size condition, it was hypothesized that larger numbers could be detected more easily in peripheral vision. This prediction was supported by the data. Fixations elicited to the easier, larger numbers were less numerous, with longer saccadic amplitudes. Lower spatial frequency information can be used to guide the eyes to bigger numbers further from the fovea, causing fewer local shifts of attention within the scanpath. Larger font sizes did not seem to increase the saccade landing variability, indicating that saccades are directed toward the center of gravity of a number rather than specific parts of it. Neverthess, this is a good example from real data of where a multi-pronged approach is needed. Neither the saccadic amplitude measure, nor the scanpath similarity results which follow allow for strong claims to be made about saccadic targeting.

Set size alternatively, reveals a pattern of more numerous fixations and shorter saccades, owing to the presence of distractors. The number of fixations overall were highest here, whilst saccadic amplitudes were shortest. Inspection of the distribution of fixations around target locations does not support more precise saccade targeting as the number of distractors increase. Noise level was intermediate in these data, indicating a general perturbation of visual search when the number targets are less visible.

It is notable that no significant effects on fixation duration were found. One possibility to account for this, is that task-difficulty-related differences in fixation duration are often explained in terms of cognitive processing effort—the word frequency effect (Rayner & Raney, 1996), the effect of informational load (Gould, 1973), and the difficulty of extracting information from a display (Goldberg & Kotval, 1999), for example, all account for fixation duration increases with harder stimuli in terms of mental effort, not in terms of physical properties of the stimuli themselves. With the present study however, difficulty is manipulated perceptually, not by the individual numbers under inspection. Nevertheless, one still might plausibly expect fixation duration differences to be present because under certain conditions purely perceptual factors such as luminance and spatial frequency, can influence fixation times (Loftus, 1985; Mannan et al., 1995). Greater quantitative detail about fixation durations is an avenue where the strength of a Multidimensional scanpath comparison can show; because MultiMatch compares fixation times pairwise between fixations in the aligned vector sequences, this has the potential to reveal stable commonalities in fixation duration at specific points along the sequence paths. Given wide variances in the distributions of fixation times presenting no significant effects with traditional analyses, the duration dimension of MultiMatch can identify similarities in fixation times where the order and spatial properties of fixations are considered at the same time.

This is a good point to turn to the scanpath similarity results obtained with MultiMatch, where we will return to the issue of fixation durations, and shed further light upon the general oculomotor data in the context of a more versatile multidimensional analysis of the omnibus eye movement sequence.

3.3. Between-subjects scanpath similarity

So far we have concentrated only on general task performance and accompanying oculomotor data. How do these results fit in the context of scanpath similarity? It is evident that there is scope for eye movement sequence variability in the data presented so far. Do scanpaths become less similar as task difficulty increases, and if so, in which dimensions? To address this question and other potential outcomes, we present the scanpath similarities produced by MultiMatch at each difficulty level of our three conditions (Figures 11–13).

3.3.1. Scanpath similarity results

Each scanpath for one participant in a given trial was compared to the scanpaths of all other participants for that trial. As there were ten trials for each difficulty level, this equates to 1710 pairwise comparisons per bar in the below figures ($n = 19, k = 2$) : $n!/((nk)!k!)$. MultiMatch compared (simplified) fixation-saccade sequences directly.

The similarity data was analysed using linear mixed effects models in R through the `lme4` package Bates et al. (2014) with one predictor: Difficulty Level. Participant variation was added as a random effect. Difficulty level was coded as an ordinal variable (with 5 levels: difficulty 1–5). Similarity data were logit-transformed to better approximate a normal distribution. All plots and statistical analyses were done with transformed data (though analysis with the untransformed data produced the same statistical effects). It should be taken into account that for this analysis we are primarily interested in the effect of difficulty level for each dimension separately. Comparisons between the dimensions are not viable since each dimension has a different true zero—that is, the baseline similarity produced for comparisons between two random scanpaths differs considerably between the dimensions (see Dewhurst et al., 2012, p. 14 & 15), making statistical comparison between them invalid.

In the font size condition (figure 11), the linear mixed effect model revealed significant effects (positive slope) for the vector difference dimension (slope = 4.961, std. error = 6.608, t = 7.508, <0.0001), the length dimension (slope = 4.159, std. error = 7.085, t = 5.870, p <0.0001), and the position dimension (slope = 2.095, std. error = 7.319, t = 2.862, p <0.001). p-values were calculated using the `lmerTest package`.

This indicates that three of MultiMatch's dimensions capture the predicted influence of task difficulty on scanpath similarity. Smaller numbers are harder to locate, and people produce diverging spatial eye movements sequences when trying to fixate them in order. The shape (Vector difference), simplified saccadic amplitudes (Length), and overall spatial locus of fixations (Position) within a scanpath become more alike when the numbers are easier to locate. Note that the slopes from the linear mixed effect models do not map directly to those in figure 11, since random effects are not accounted for in the figure.

[INSERT FIGURE 11 HERE]

In the set size condition (figure 12) the effects were comparable, but with some notable exceptions. This time the vector difference dimension did not reveal a statistically significant result, whereas direction (angle) did (slope = -3.257, std. error = 7.072, t = -4.606, p <0.0001). Again task difficulty was evident in the length (slope = -2.616, std. error = 5.821 , t = -4.494, p <0.0001), and the position dimensions (slope = -3.646, std. error = 5.762, t = -6.328, p <0.0001). The slope is now negative, reflecting decreased similarity at larger set sizes. As with the behavioural and oculomotor results, the x-axis is essentially inverted from the font size condition, where larger numbers were easier, but the influence of task difficulty remains linear. A significant result was also expressed in the duration dimension, *but this time in the opposite direction from all the results reported so far*. Greater similarity in fixation durations being associated with more difficult set sizes (slope = 2.897, std. error = 5.574, t = 5.197, p <0.0001). The potential reasons for this we will return to in the discussion, whilst meanwhile directing the reader to the fact that this result is not indicative of participants having longer fixation durations at larger set sizes, as might be expected due to greater e.g. crowding in this condition (Hooge & Erkelens, 1996; Vlaskamp & Hooge, 2006)(see figure 7).

[INSERT FIGURE 12 HERE]

Finally, the noise condition (figure 13) paralleled the previous analyses, oncemore similarity remaining relatively constant across difficulty level for the vector difference dimension, but being sensitive to the manipulation for the angle dimension (slope = -1.566, std. error = 3.273, t = -4.785, p <0.0001), the length dimension (slope = -1.134, std. error = 2.502, t = -4.531, p <0.0001), and the position dimension (slope = -2.270, std. error = 2.648, t = -8.573, p <0.0001). Moreover, as with set size, the effect was reversed for the duration dimension (slope = 1.233, std. error = 2.130, t = 5.790, p <0.0001).

The general pattern in these data support the hypothesis that scanpaths diverge, becoming less similar with increasing perceptual difficulty of the task.

[INSERT FIGURE 13 HERE]

4. General Discussion

It is encouraging with respect to our previous paper (Dewhurst et al., 2012), that MultiMatch can cope with a less constrained and more variable visuoperceptual task, comparing the multiple scanpaths produced, and in many cases identifying the hypothesized decrease in scanpath similarity scores with increasing task difficulty. For a visual search-like task where there is no implicit advantage for our method from the outset, and the type and nature of similarity was not known in advance, the results obtained from MultiMatch shed further light upon the oculomotor mechanisms underlying search performance than can be known from basic eye movement statistics alone. However, it is also worth reiterating here that what MultiMatch provides is extra, not better information.

4.1. From behavioural data to oculomotor statistics

The behavioural data showed the expected rise in search times in each condition as the conspicuity or visibility of the number stimuli became weaker. Coupled with slight speed-accuracy trade-offs in the font size and set size conditions, reflecting somewhat speeded responses when these tasks were at their easier difficulty levels, and the need for more deliberate, slower search when the task became harder, these data paint a common picture of scanning behaviour when the searched for item becomes more difficult to locate. To unpack these eye movement trends, we monitored three common parameters of search: number of fixations, fixation duration, and saccadic amplitude. The first and the third of these measures growing in frequency and declining in amplitude, respectively towards the harder end of the difficulty scale. This was the hypothesised pattern of search behaviour, and is typical of similar search tasks in which eye movements are recorded (Zelinsky & Sheinberg, 1997).

4.2. Multi-dimensional scanpath similarity

By themselves results such as this are revealing about how we inspect, identify, deselect and progress in search. But by adding the multiple dimensions of scanpath comparison as well, we gain more insight into the process of search as a whole. The most stable dimensions which MultiMatch consistently identified similarity reductions in were Length and Position. In all conditions (font size, set size, and noise level) participants fixated comparable locations with highly similar saccadic amplitudes, and note that this is not merely spatial similarity, since MultiMatch attempts to retain the order of the scanning sequence. In short, observers inspect similar positions in a similar order with similar saccadic targetting—and this group style tendency becomes less pronounced when there

are greater demands on the visuoperceptual system. The more spatial, or shape related dimensions of MultiMatch (Vector and Angle) differed slightly between conditions. Declining Vector similarity scores with greater difficulty level were found for different font sizes; this was not the case for the other two conditions, where the angle dimension instead was sensitive to the manipulation. This may be owing to the larger sizes of the numbers when the task is easier.

4.2.1. Implications for fixation duration

Interestingly, the Duration dimension showed *higher* similarity scores in the set size and noise level conditions when the task became harder. It was pointed out in the results section that this is not simply due to extended fixation durations overall, since the general oculomotor data show no gain in fixation times, remaining constant (∼200ms) irrespective of condition and difficulty. Its is crucial to highlight that the strength of MultiMatch is made apparent here in considering positions in the sequence of the scanpath as a whole. Because the position dimension maintains relatively high similarity scores, consistently finds an effect of task difficulty, and fixation sequence order is broadly retained, we can be confident that the high similarities in the duration dimension are not simply randomly distributed in space. Rather, people elicit more similar fixation durations *in more similar positions* at around the same time in their search. This gives more information about fixation times than basic fixation duration statistics, even if they are broken down across AOIs (because MultiMatch compares fixation times pairwise between fixations in the aligned vector sequences, which may not adhere to strict AOI boundaries). In the not uncommon case of gathering data with wide variances in the distribution of fixation times, presenting no significant differences between conditions, MultiMatch can help the researcher identify stable commonalities within those distributions at specific points along the scanpath's route. This sort of analysis could also be accomplished by binning fixations into different AOIs, as well as splitting into different time periods or comparing fixations in sequence. However, such an analysis would require numerous comparisons and arbitrary decisions about the bins involved (see Orquin et al. (2016) for a good discussion of the issues surrounding AOI selection).

Nevertheless, we must consider the underlying reasons behind the increase in Duration similarity when the set and noise level conditions become harder. Perhaps participants impose a temporal upper limit on fixation times (cf. Henderson, 1992; Henderson & Pierce, 2008) to offset the effects of crowding and lower conspicuity (Hooge & Erkelens, 1996; Vlaskamp & Hooge, 2006) in these more visually demanding cases, so as to ensure maximum coverage of the area to be inspected, without a search time cost. This is an interesting avenue for further study in itself, because it generates the prediction that observers can strategically tune the efficiency of information extraction within fixations in accordance with task demands. One could then ask, is this automated, or by voluntary adaptation of the saccadic system? Where visibility is in general better, in the font size condition, such adaptation may not be necessary. Furthermore, it is notable also that the variability in fixation durations tends to shrink at higher difficulty levels in the set size and noise conditions (Figure 7). This adheres to the argument that fixation durations become more alike, drawn from MultiMatch's Duration dimension similarity results. However, the reader should be aware that this convergence of fixation times does not necessarily indicate that the absolute value of fixation durations is approximately the same between participants, since the duration dimension still returns much lower similarity scores than the other dimensions of MultiMatch—just below 0.6 on average (see figures 11–13). Strictly speaking however, comparison between dimensions is not viable since they have different baselines when comparing random scanpaths (see Dewhurst et al., 2012, p. 14 & 15). What we can say, is that fixation durations at matched points along the scanpath begin to

resemble each other, in relative terms, when the task is harder compared to when it is easier.

We should not be surprised that the Duration dimension detects lower similarity in fixation times overall. This fits very well with the known idiosyncrasies in fixation duration seen in the literature (e.g. Andrews & Coppola, 1999; Rayner et al., 2007). It is therefore quite parsimonious that people exhibit less similar/more different fixation durations between subjects, even when inspecting the same stimulus array. Moreover, as well as producing an expected relative lack of similarity *between-subjects* on the duration dimension, we have elsewhere shown that MultiMatch is capable of producing the converse for *within-subjects* comparisons, where fixation durations *should* be equivalent. Foulsham et al. (2012) have shown with MultiMatch, that in picture viewing, individuals' fixation durations are more similar to their own, even when viewing different images. This source of idiosyncratic similarity remains more powerful even between images, than the similarity observed for different people viewing the same image.

It is reassuring that when analyzing the whole scanpath representation, results compatible with the fixation duration literature come out.

4.3. Spatial extent and the use of AOI's

Here the strengths and weaknesses of MultiMatch are exposed. MultiMatch does not require the definition of AOIs as regions in space, and although scanpaths are simplified via experimenter-defined thresholds, this means that more of the original scanpath representation remains. Combined with the five dimensions then, the potential for revealing similarity is very high. But it is also evident that this may be too much, leading to high similarity in many many cases therefore increasing the chances of Type 1 errors. Plus, the similarity scores are difficult to interpret with the layers of first simplification, then alignment before the different dimensions of similarity can be calculated, and finally statistically analysed. Reliance on AOIs with other scanpath comparison tools such as ScanMatch (Cristino et al., 2010) for instance, means that although quantization errors can be made (see Dewhurst et al., 2012; Anderson et al., 2015), potentially more meaningful differences between scanpaths are detectable.

The issue of AOIs also plays a role when we consider the similarity results for the font size condition. If we turn to the area occupied by the larger numbers, although larger numbers are easier to locate, therefore in some respects, such as shape (vector), giving rise to similar scanpaths, there is also more variability in saccade landing position which will still count as a correctly targeted saccade for the LCS operationalisation of accuracy. With bigger (easier) numbers this means that participants can hit the target AOIs sufficiently well, even in exactly the right order, and still produce subtly different scanpaths, simply because the spatial area is extended. This could have the effect of yielding less similar scanpaths for larger numbers than would otherwise be the case if the numbers were fixated exactly in the centre of their respective AOIs. Thus, the effect of task difficulty might be weakened in this condition, perhaps explaining why this condition returns difficulty-induced similarity reductions in the Vector difference dimension but not the Angle dimension, while the opposite is true for the other two conditions. Nevertheless, the heatmaps we produced to address this issue in saccadic targeting make this explanation less likely, and it is beyond the scopt of this paper to dig into saccadic targeting more critically.

There are also methodological issues regarding the choice of AOI size that could influence the results. In this work, a margin of two degrees were added to a square encompassing each number. Instead of a fixed value, the AOI margin could be selected based on the size of the number or the calibration accuracy of the participant. Systematically varying the size of the AOI margin and

see how that influences the results could by itself potentially provide information about the eye movements behavior.

These factors should not necessarily be viewed as strengths nor weaknesses of MultiMatch per se. The data and possible explanations for it are given simply to assist the user in interpreting output from MultiMatch. We hope this is timely given the growing popularity of MultiMatch and scanpath measures generally in eye movement research (Foerster & Schneider, 2013; Anderson et al., 2015; French et al., 2016; Kübler et al., 2016; Wilson et al., 2018). At the same time, it should be stressed that MultiMatch is not a universal measure suitable for all research questions where scanpaths are recorded; different research question may require a different measure of combination of measures.

4.4. Summary & Conclusions

The present study investigated the effects of perceptually induced task difficulty on eye movements, and in particular on the between-observer similarity in scanpaths computed using our MultiMatch method. The results showed a range of effects which varied according to the manipulation of difficulty. These data reveal new insights into human visual search performance, as well as providing more information about MultiMatch as an analytical tool in eye movement research.

In general, more difficult number displays led to an increased number of fixations and shorter saccades, which is consistent with findings from prior research, and a generally prolonged scanpath. Moreover, participant scanpaths were less similar to each other in more difficult trials. This finding arises because of differences in a number of dimensions describing the spatial extent of the scanpath as well as the duration of fixations. In future work, we could investigate other interactions between these dimensions. For example, the relationship between fixation duration and saccade amplitude over time (Unema et al., 2005) could also be defined in terms of changing sequential organisation of eye events (i.e., a scanpath). We would therefore expect higher MultiMatch similarity between scanpaths from the same viewing mode (e.g., 'early' or ambient viewing) than between different modes.

In sum, we identified that as a task gets harder, participants change their eye movements. As in more complex and applied tasks, it can be tempting to summarise this as resulting in a different scanpath. Our results show, however, that individuals do not 'fail' in a uniform way when difficulty increases. Instead, they become more idiosyncratic and less similar to each other. While the stimulus-driven causes of changes in the eye movement record (such as the crowding from increased distractors, or the decreased perceptibility in peripheral vision) have been previously examined, such changes pose a challenge for measuring scanpath similarity. This paper argues for a multidimensional approach in linking differences in average oculomotor measures to omnibus, but sequential, changes in scanpath shape, length and the duration of fixations of which the scanpath consists.

5. Acknowledgements

This work was supported by a post-doctoral research grant from the Swedish Research Council awarded to Richard Dewhurst (grant no. 435-2010-849), and the Linnaeus center for Thinking in Time: Cognition, Communication and Learning (CCL) at Lund University, which is also funded by the Swedish Research Council (grant no. 349-2007-8695).

AppendixA. Influence of MultiMatch thresholds on similarity scores

To investigate the sensitivity of thresholds on MultiMatch scores, the amplitude and direction threshold were systematically varied between 5 to 15% of the screen width in steps of 2 (amplitude) and 30 to 60 degrees in steps of 6 (direction threshold). The simulation was run with scanpaths of a fixed length ($n = 10$) and randomly drawn positions. Unsurprisingly, there are differences in some of the dimensions due to changes in thresholds. However, using values close to the 'default' ones used in this paper introduce only small changes in similarity.

[INSERT FIGURE A.14 HERE]

Anderson, N. C., Anderson, F., Kingstone, A., & Bischof, W. F. (2015). A comparison of scanpath comparison methods. *Behavior research methods*, *47*, 1377–1392.

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*, 2947–2953. doi:DOI: 10.1016/S0042-6989(99)00019-X.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4. URL: `http://arxiv.org/abs/1406.5823` arXiv e-print; submitted to *Journal of Statistical Software*.

Brainard, D. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

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

Dewhurst, R., & Crundall, D. (2008). Training eye movements: Can training people where to look hinder the processing of fixated objects? *Perception*, *37*, 1729–1744.

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*, (pp. 1–22).

Foerster, R. M., & Schneider, W. X. (2013). Functionally sequenced scanpath similarity method (funcsim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units. *Journal of Eye Movement Research*, *6(5)*, 1–22.

Foulsham, T., Dewhurst, R., Nyström, M., Jarodzka, H., Johansson, R., Underwood, G., & Holmqvist, K. (2012). Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Jounral of Eye Movement Research*, *5 (4)*, 1–14.

Foulsham, T., & Kingstone, A. (2010). Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. *Vision Research*, *50*, 779–795.

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*.

French, R. M., Glady, Y., & Thibaut, J.-P. (2016). An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making. *Behavior Research Methods*, (pp. 1–12).

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and construct. *International Journal of Industrial Ergonomics*, *24*, 631–645.

Gould, J. D. (1973). Eye movements during visual search and memory search. *Journal of Experimental Psychology*, *98*, 184.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on raven's advanced progressive matrices. *Journal of vision*, *11*, 1–11.

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition* (pp. 260–283). Springer.

Henderson, J. M., & Pierce, G. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin & Review*, *15*, 566–573.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 210–228.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R. C., Jarodzka, H., & van der Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hooge, I., & Erkelens, C. (1996). Control of fixation duration in a simple search task. *Attention, Perception, & Psychophysics*, *58*, 969–976.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 211–218). ACM.

Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1289.

Kübler, T. C., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2016). Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods*, (pp. 1–17).

Loftus, G. R. (1985). Picture perception: Effects of luminance on available information and information-extraction rate. *Journal of Experimental Psychology: General*, *114*, 342.

Madsen, A., Larson, A., Loschky, L., & Rebello, N. S. (2012). Using scanmatch scores to understand differences in eye movements between correct and incorrect solvers on physics problems. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 193–196). ACM.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*, 109–127.

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

Mathworks (). Box plot. URL: `http://www.mathworks.se/help/stats/boxplot.html` accessed 15/07/12.

Ni, J., Jiang, H., Jin, Y., Chen, N., Wang, J., Wang, Z., Luo, Y., Ma, Y., & Hu, X. (2011). Dissociable modulation of overt visual attention in valence and arousal revealed by topology of scan path. *PloS one*, *6*, e18262.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data. *Behavior Research Methods*, *42*, 188–204.

Orquin, J. L., Ashby, N. J., & Clarke, A. D. (2016). Areas of interest as a signal detection problem in behavioral eye-tracking research. *Journal of Behavioral Decision Making*, *29*, 103–115.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). Investigating the visual span in comparative search: The effects of task difficulty and divided attention. *Cognition*, *81*, B57–B67.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, *62*, 1457–1506.

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research*, *47*, 2714–2726.

Rayner, K., Miller, B., & Rotello, C. (2008). Eye movements when looking at print advertisments: The goal of the viewer matters. *Applied Cognitive Psychology*, *22*, 697–707.

Rayner, K., & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, *3*, 245–248.

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

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, *12*, 473–494. doi:http://dx.doi.org/10.1080/13506280444000409.

Vlaskamp, B., & Hooge, I. (2006). Crowding degrades saccadic search performance. *Vision research*, *46*, 417–425.

Wilson, K. A., Heinselman, P. L., & Kang, Z. (2018). Comparing forecaster eye movements during the warning decision process. *Weather and Forecasting*, *33*, 501–521.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23(1)*, 244–262.

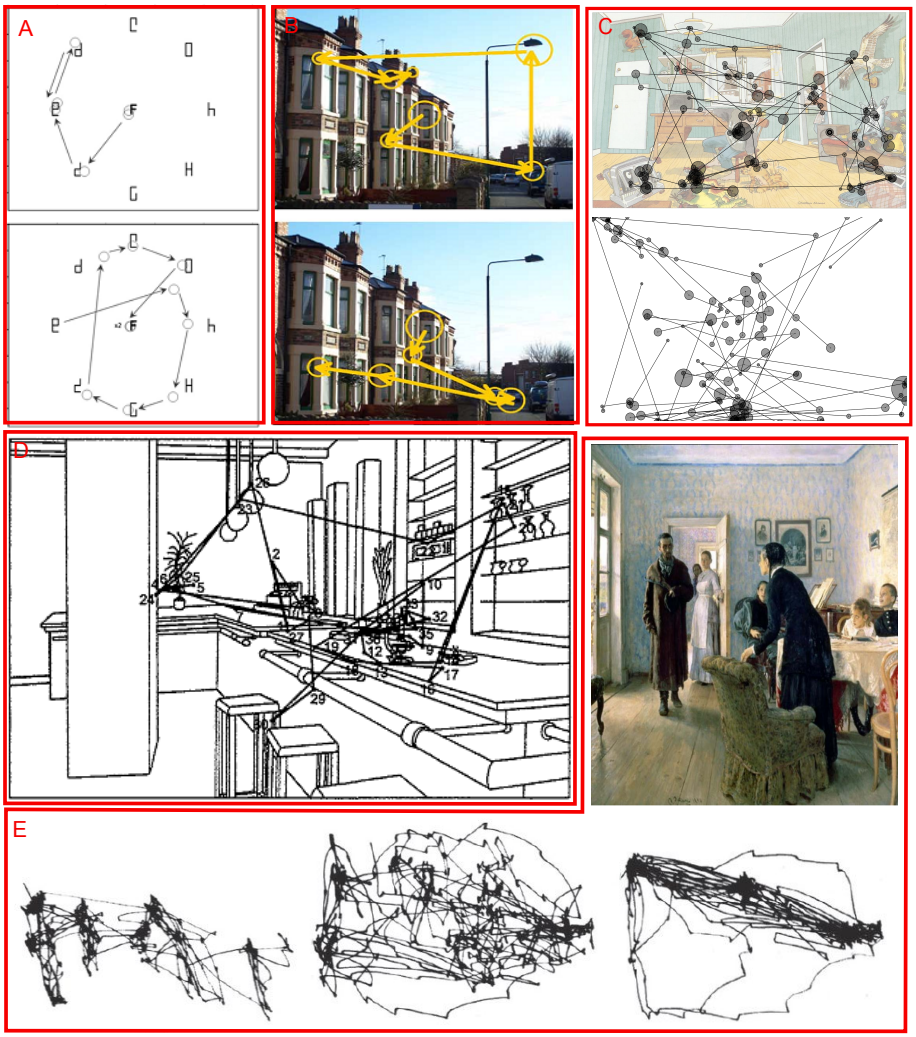


Figure 1: Scanpath visualizations from participants: **A**. Inspecting a central letter then searching amongst peripheral items (Dewhurst et. al., Training Eye Movements: Can Training People Where to Look Hinder the Processing of Fixated Objects? Perception (37 11), pp. 1738. ©[2008] SAGE publications. Reprinted by permission of SAGE Publications.); **B**. Encoding and recognising an image (Reprinted with permission from 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), p.10. ©2008 ARVO.); **C**. Encoding and imagining a picture (original data published in Johansson et. al., 2012); **D**. Inspecting a bar scene (Henderson et. al. The effects of semantic consistency on eye movements during complex scene viewing, Journal of experimental psychology: Human perception and performance, 25(1), p. 214 , 1999. Published by The American Psychological Association (APA), and reprinted with permission.); **E**. Looking at a picture for different purposes (Adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH. Eye movements and vision, p. 172, by Alfred Yarbus. ©Springer 1967. Upper panel: "The Unexpected Visitor". Oil on canvas painting by Ilya Repin, 1884-88. Source: Courtesy of www.ilyarepin.org).

Figure 2: Examples of stimulus displays and tasks (not to scale). Larger font size (left), more distractors (middle), and greater background noise (right). These factors were manipulated independently. The task is to look at the numbers in increasing order, i.e. 1,2,3,4,5. All numbers were displayed simultaneously within one trial.

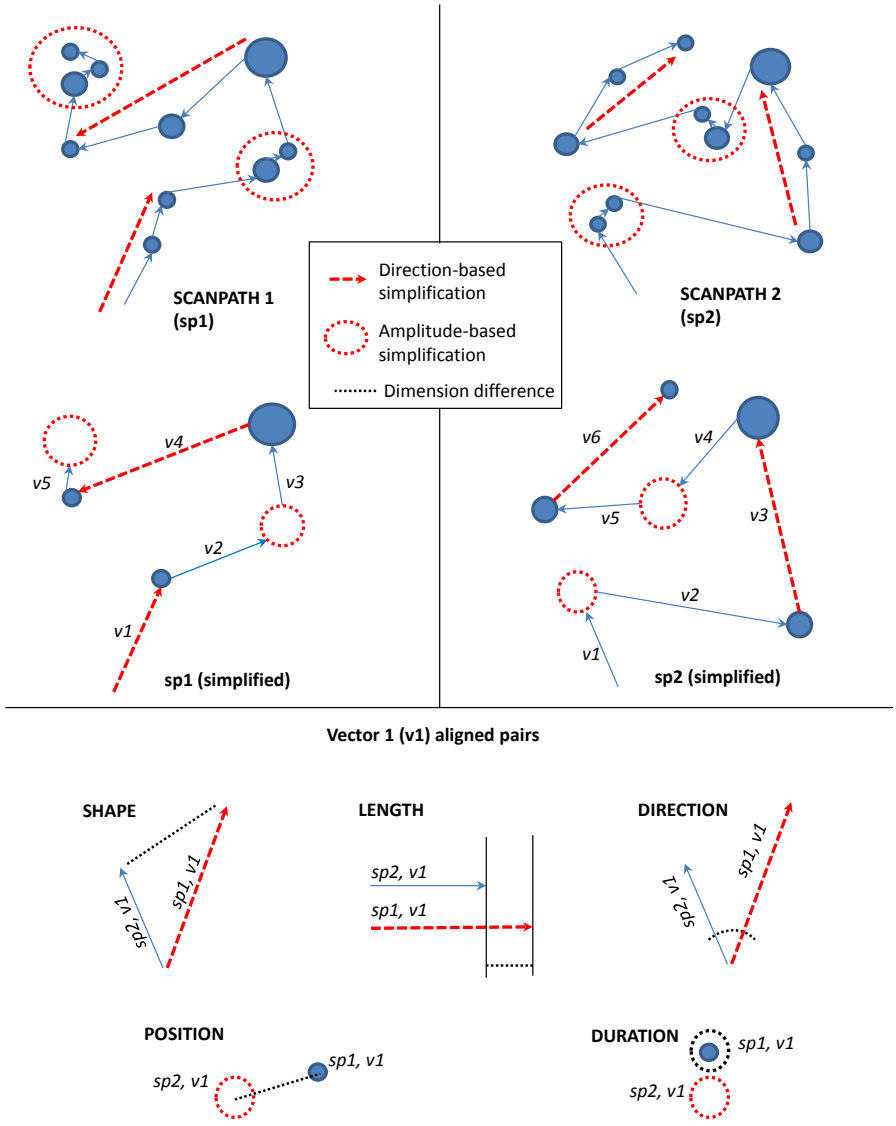
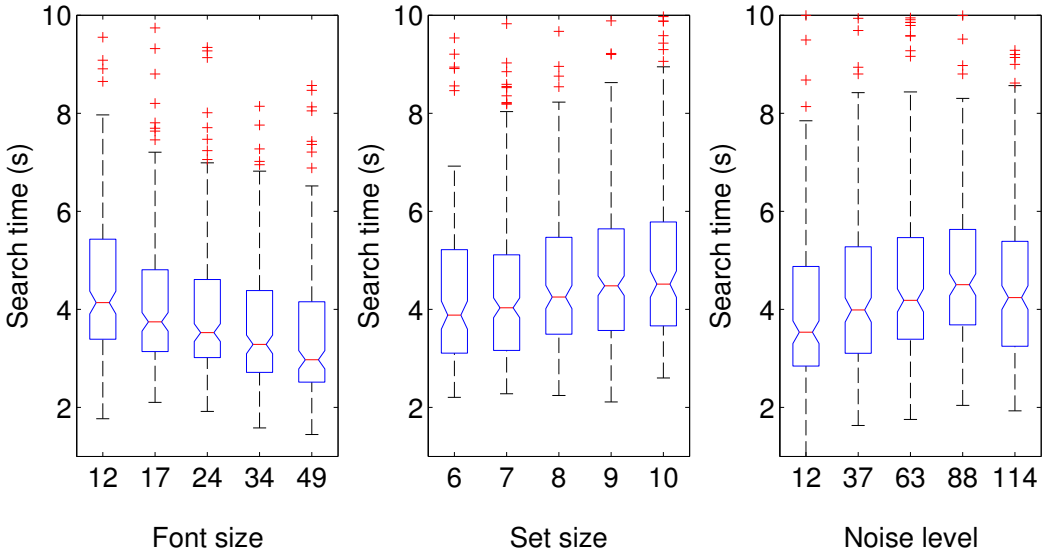
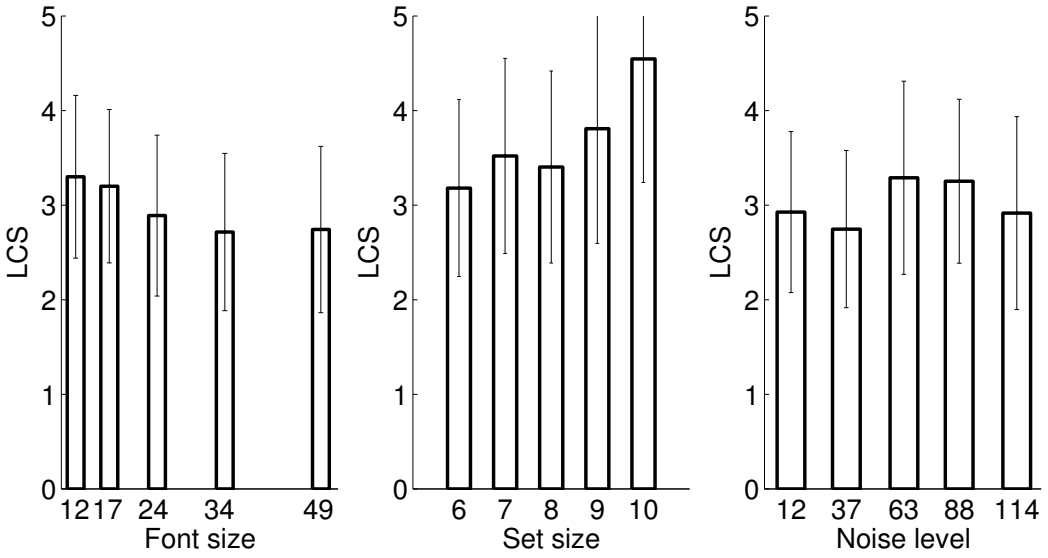


Figure 3: Two hypothetical scanpaths. Blue arrows depict saccades, blue circles fixations (larger circle = longer fixation duration). In MultiMatch, first each scanpath is simplified according to the direction (dashed red arrows) and amplitude (dashed red circles) of saccades. When subsequent saccades continue within an angle of 45° of the preceding saccade, these are collapsed into one vector—Direction-based simplification. When following saccades are smaller than 10% of the screen diagonal they are likewise grouped into a single vector—Amplitude-based simplification. These are not hard thresholds, but have proven to be good based on our testing of the algorithm; adjusting them does not substantially affect the results described here. After simplification, the next step is to align the vectors of each scanpath using the Dijkstra algorithm (1959). The lower panel illustrates dimension differences for the first pair of aligned vectors; v1 between simplified scanpaths sp1 and sp2. The numeric difference between each dimension is illustrated with a dotted black line for each dimension separately.



(a) (b) (c)

Figure 4: Search time as a function of difficulty level for each condition. Boxplots show medians dividing each box. The edges of the box are the 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Non-overlapping notches indicate significantly different medians at the 5% level. These figures are for visualisation and the analysis sections in the text are carried out separately.



(a) (b) (c)

Figure 5: LCS as a function of difficulty level for each condition. Bars represent mean values and error bars standard deviation.

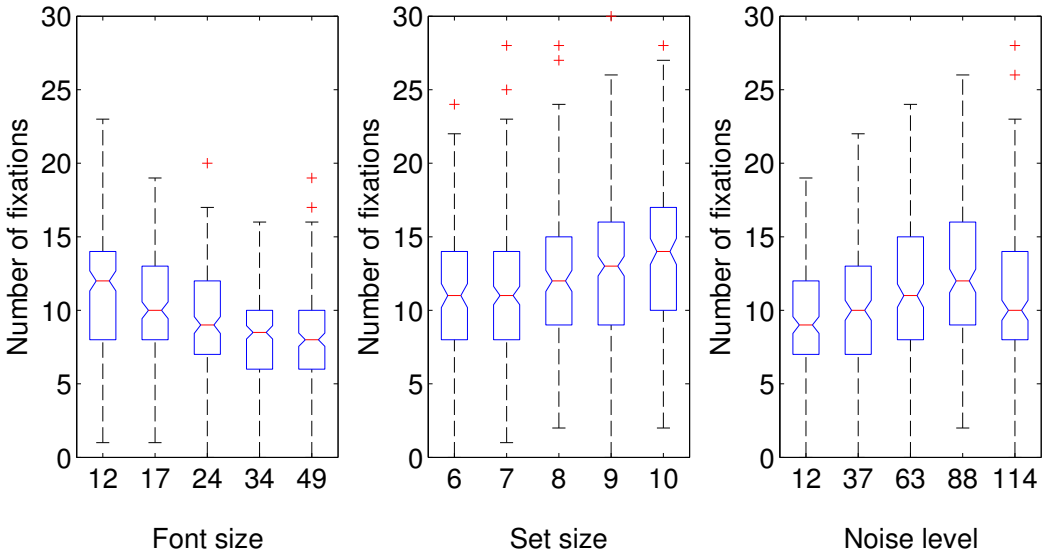


Figure 6: Number of fixations as a function of difficulty level for each condition. Boxplots show medians dividing each box. The edges of the box are the 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Non-overlapping notches indicate significantly different medians at the 5% level. These figures are for visualisation and the analysis sections in the text are carried out separately.

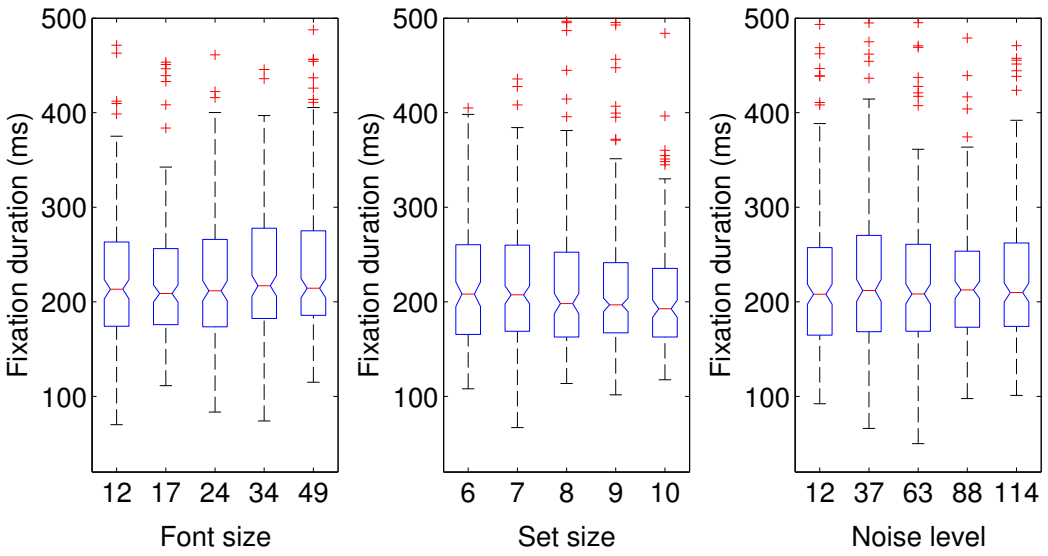


Figure 7: Fixation duration as a function of difficulty level for each condition. Boxplots show medians dividing each box. The edges of the box are the 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Non-overlapping notches indicate significantly different medians at the 5% level. These figures are for visualisation and the analysis sections in the text are carried out separately.

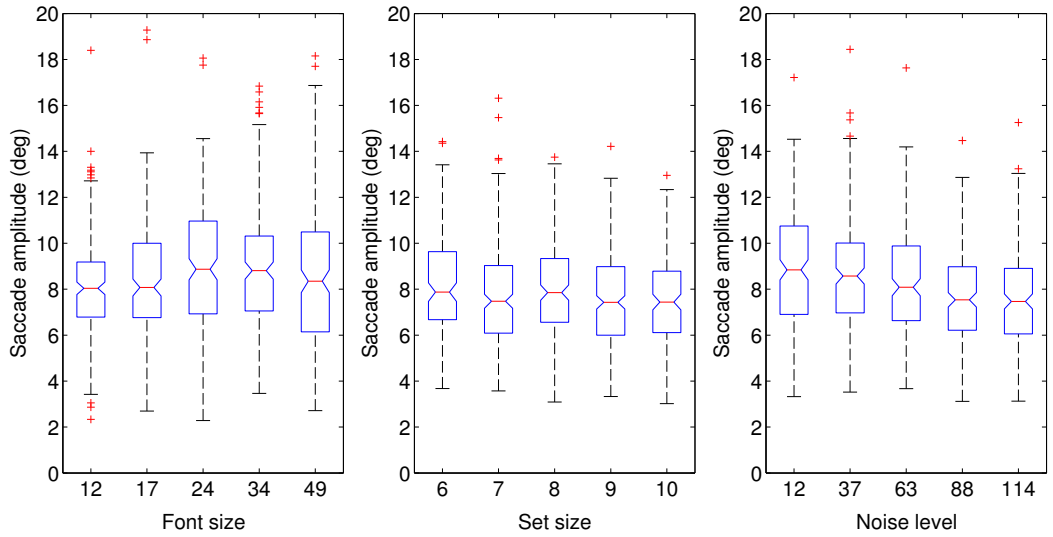


Figure 8: Saccade amplitudes as a function of difficulty level for each condition. Boxplots show medians dividing each box. The edges of the box are the 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Non-overlapping notches indicate significantly different medians at the 5% level. These figures are for visualisation and the analysis sections in the text are carried out separately.

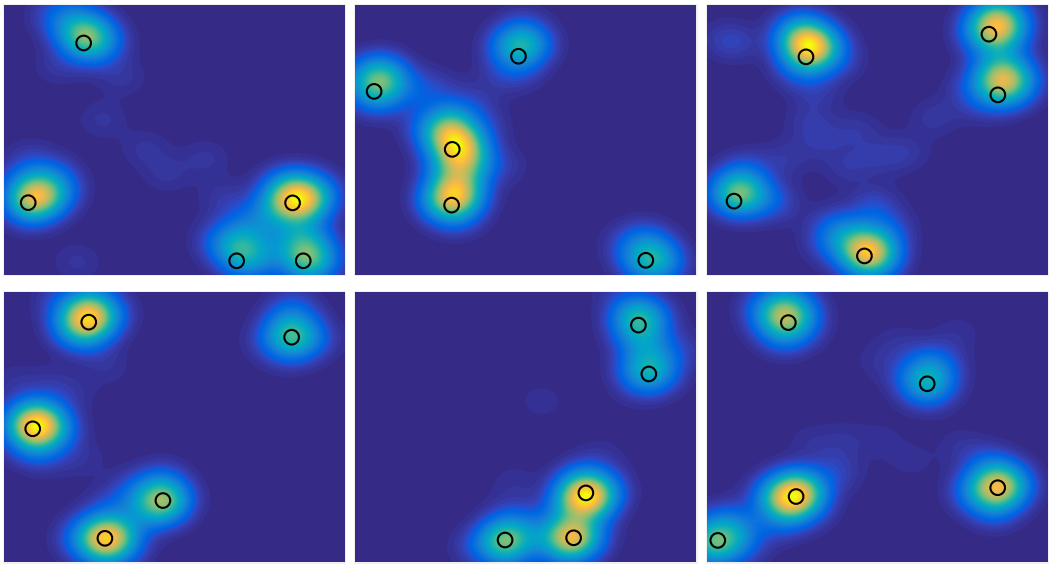


Figure 9: Heat maps illustrating how fixation distributions change when the font size is large (top row) and small (bottom row). Circles represent locations of the target numbers.

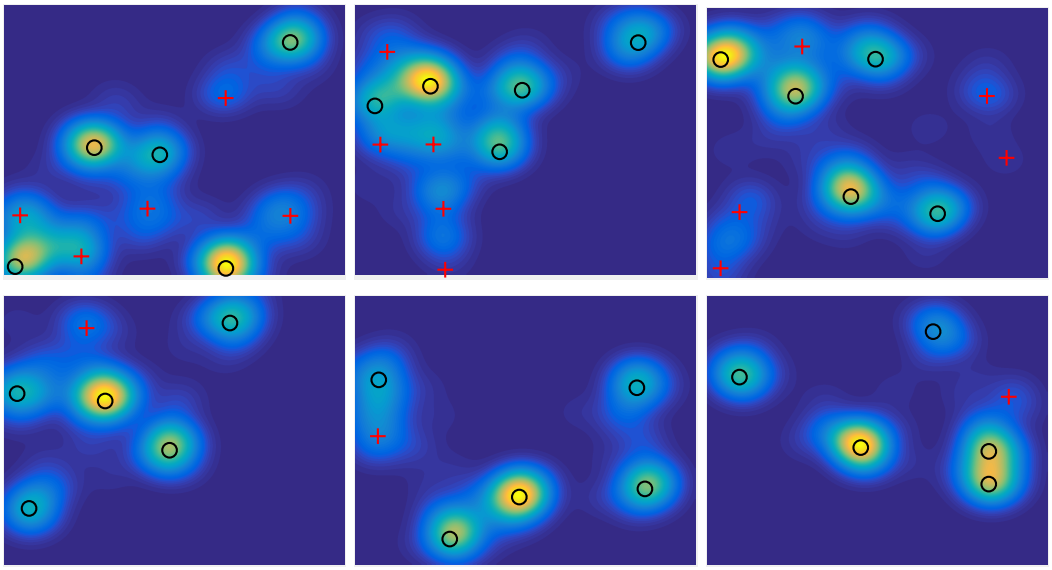


Figure 10: Heat maps illustrating how fixation distributions change when the set size is large ($n = 10$, top row) and small ($n = 6$, bottom row). Black circles represent locations of the target numbers and red plus signs (+) represent distractor locations.

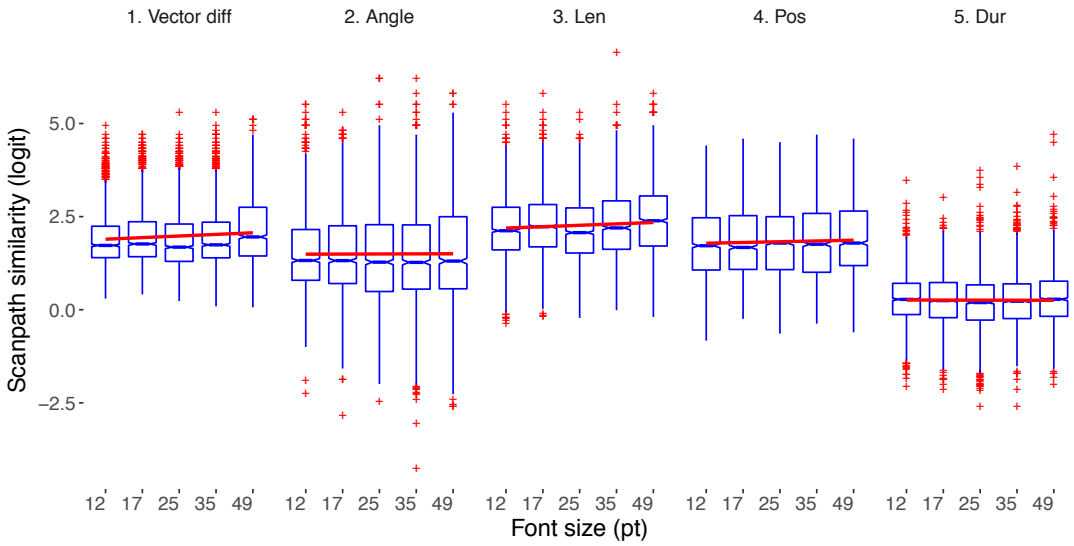


Figure 11: Basic similarity results in the font size condition for each of MM's dimensions at all difficulty levels (font size 12 = difficult, to font size 49 = easy). The upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles). Whiskers extend 1.5 times the inter quartile range, and points outside the whiskers represent outliers. Notice that the scanpath similarity results have been logit-transformed. The trend lines represent a linear fit to the data.

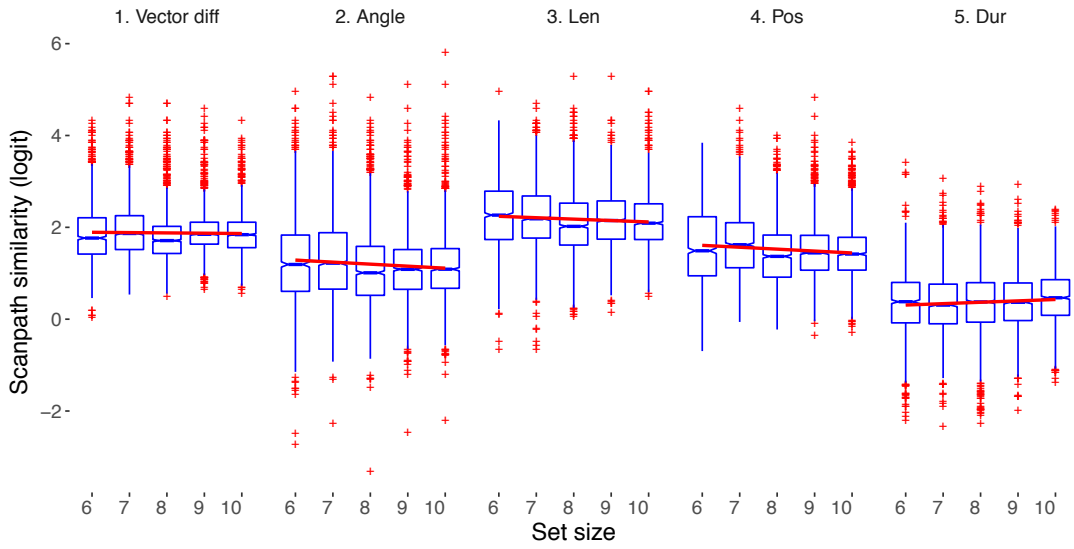


Figure 12: Basic similarity results in the set size condition for each of MM's dimensions at all difficulty levels (numbers 1–6 = easy, to numbers 1–10 = difficult). The upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles). Whiskers extend 1.5 times the inter quartile range, and points outside the whiskers represent outliers. Notice that the scanpath similarity results have been logit-transformed. The trend lines represent a linear fit to the data.

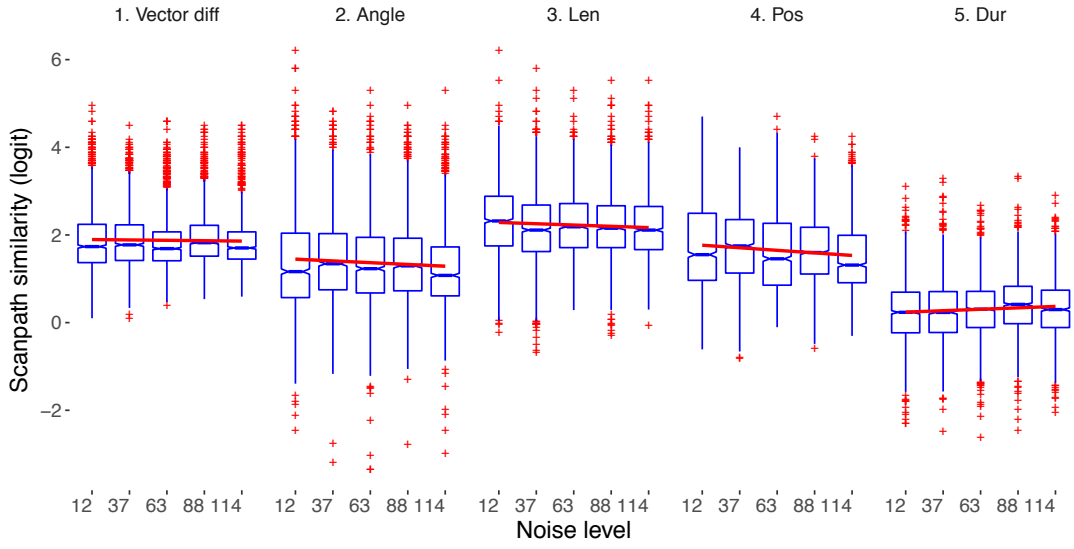


Figure 13: Basic similarity results in the noise level condition for each of MM's dimensions at all difficulty levels (noise level 12 = easy, to noise level 114 = difficult).The upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles). Whiskers extend 1.5 times the inter quartile range, and points outside the whiskers represent outliers. Notice that the scanpath similarity results have been logit-transformed. The trend lines represent a linear fit to the data.

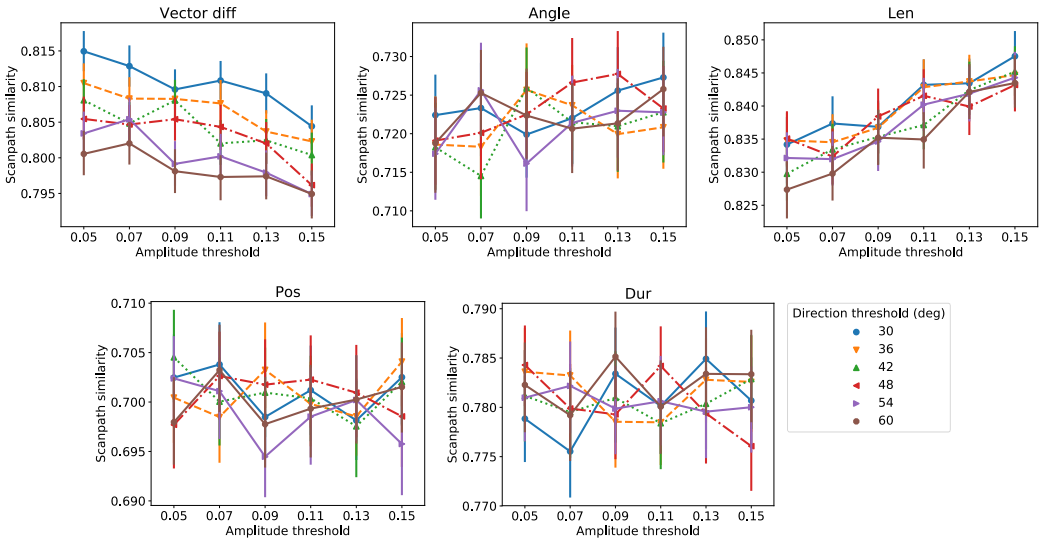


Figure A.14: Influence of amplitude and direction thresholds on MultiMatch scanpath similarity. Similarity values were calculated for scanpaths of length $n = 10$ with randomly drawn positions.