What do Scanpaths Tell Us About Cognitive Processes? An Investigation in a Problem Solving Domain

By:

Samantha STRANC

A thesis submitted to the Faculty of Graduate and Post Doctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Cognitive Science

Samantha STRANC
in

Master of Cognitive Science program

Carleton University,
Ottawa, Ontario
Samantha STRANC

Abstract

Scanpaths are the specific sequence of fixations elicited by someone when viewing a scene or object. Not only do they illustrate which areas of the scene that were fixated on, they also capture the viewer's change in attention overtime. Dynamics of visual attention contain movement patterns that are not otherwise measured with simple fixation methods of eye data analysis. In this thesis, we employ MultiMatch, a scanpath analysis method that provides quantitative measure of similarity between two scanpaths, to examine visual patterns for two different data sets. These data sets came from studies that presented students with math problems and varied instructional material to manipulate student's solving strategies. We apply an analysis method corresponding to grouping scanpaths across and within-conditions to determine whether the MultiMatch analysis method can distinguish between the instructional material presentation formats. We further our analysis by providing initial interpretation guidelines through a brief scanpath simulation model. Results demonstrate a difference in viewing pattern use between conditions in the original studies, which were designed to elicit different solving strategies.

Acknowledgements

My time at the department of Cognitive Science at Carleton University has at once passed in a blur and also felt it could last a lifetime. Given the people; the professors and fellow students alike, I would consider myself lucky for any amount of additional time spent with the department. The duration of my degree has been a time of learning, involvement and growth. Thank you.

To all my family and friends, thank you for your continued help and showing up with words (and food) of encouragement. I could not have done it without your making time and space for my vague questions. Thanks especially to my brother, Colin, for coding and moral support and to my partner, Jay, for Latex and additional moral support.

To Dr. Kasia Muldner, who has been the most patient of all - thank you for helping make my analysis dreams a reality, for providing endless support, motivation and opportunities along the way.

Contents

List of Figures

List of Tables

1 Introduction

A scanpath is the sequence of fixations that occur while a viewer looks at some target, for instance an image or a classroom scene (Noton and Stark, 1971). Given that scanpaths encode visual attention patterns, they have the potential to inform on cognitive strategies and/or processes. To illustrate, we will provide a hypothetical example. Suppose a student is looking at an algebra program that is presented alongside a similar worked-out example, which provides the solution. If we recorded the student's eye movements while they were trying to solve the problem, we would observe the student's gaze fixating on the problem and then switching to the equivalent spot in the worked-out example; this switching back and forth repeats until the solution is generated. Although we don't have access to what this student is thinking, studies that examined student reasoning patterns found that this pattern of switching back and forth between the problem and example is indicative of generating the problem solution through copying the example solution (VanLehn, 1996).

We will provide another, alternative scenario. Suppose a different student is looking at the same problem and example pair, but exhibiting considerably different eye movement patterns. The second student's gaze focuses on the problem description for some time before switching to the worked-out example – here the student fixates on the example solutions lines for some time, occasionally fixating on previous lines and/or the example description. Eventually, the student's gaze returns back to the problem and stays focused on that area. While we do not have access to this student's thought process either, this pattern of eye movements is indicative of self-explanation (Chi et al., 1989). In particular, the second student studies the example solution, infers the principles needed to solve the problem and only then moves back to the problem-solving area. These two hypothetical students fixated on similar problem and example areas but the order of those fixations was different. It was the order of fixations that indicated the strategies that students were engaging in as they problem solved.

As illustrated with these hypothetical examples, the advantage of observing scanpath patterns is that they capture not only where people look but also the order in which areas are looked at. This advantage comes with a challenge, in that scanpath analysis is more complicated than basic fixation analysis. Existing tools for scanpath analysis vary widely but typically produce a similarity score characterizing the degree of alignment between two scanpaths. In this thesis, we used the MultiMatch analysis tool (Jarodzka et al., 2010; Dewhurst et al., 2012). MultiMatch aligns and compares two scanpaths to quantify how similar they are on a scale of zero to one; the similarity score is generated for five different scanpath features, including shape, position, length, direction and duration. How to analyze and interpret the scanpath similarity data is another challenge, as there is relatively little work providing guidance on these aspects (but with notable exceptions, Dewhurst et al. (2012), Dewhurst et al. (2018), Zhou et al. (2016)). In general, more work is needed to shed light on the utility of scanpath analysis including its advantages and limitations. Moreover, to date, scanpath work has targeted domains outside of educational settings, and so less is known about their utility in these settings.

The present thesis aims to contribute to the field of scanpath investigation by analyzing two different eye-tracking data sets (Tan, Muldner, and LeFevre, 2016; Jennings and Muldner, 2020). The first data set comes from a study investigating the impact of format in basic arithmetic problems, i.e., 18 ÷ 3 = [ ] (Tan, Muldner, and LeFevre, 2016). The second data set comes from a study investigating the impact of example format for solving algebraic math problems, such as "b = ((c(a+f))÷d)-e" (Jennings and Muldner, 2020). Thus, in both studies students had to solve math problems but these problems varied in complexity from basic (study 1) to more complex (study 2). The goal for the current work is to apply scanpath analysis methods to examine whether scanpaths distinguish between instructional material format. As will be described in detail in future chapters, scanpaths were compared within the target conditions and across the conditions. We predicted the higher within-condition comparisons will produce higher similarity scores compared to across-condition comparisons, because the study conditions were designed to elicit different strategies and we expected those strategies to be reflected in scanpaths. This study is merely a first step in exploring the applicability of scanpaths as another outcome variable that can distinguish between complex cognitive processes elicited by different conditions.

Before describing the work conducted for the thesis, we present a summary of research on scanpath analysis; subsequent chapters will provide further background when the corresponding studies are described.

2 Related Work

To date, eye tracking analysis has predominantly relied on fixation data (e.g., Sharafi, Soh, and Guéhéneuc (2015); Lai et al. (2013); Mayer (2010)) - for a comprehensive review, see Henderson and 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 using 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 for expertise in diverse domains, such as chess (Reingold et al., 2001), epilepsy diagnosis (Jarodzka et al., 2012), and computer program construction (Nivala et al., 2016). While fixation data is informative, it ignores the order in which fixations occur. The chronological order is captured by the viewer's scanpath, which as defined in the introduction is a sequence of fixations, including fixation location and duration.

2.1 Scanpath Analysis: Foundations

A common approach to scanpath analysis involves comparing the similarity of scanpaths coming from different experimental conditions and/or populations, in order to determine if these factors impact visual attention patterns. This is based on the assumption that if tasks and individual differences impact cognitive processes, we would expect to see these differences reflected in visual attention patterns captured by scanpaths. Thus, scanpath similarity should vary between different groups and/or conditions. The foundation for this approach involves quantifying the similarity between pairs of scanpaths. There are many methods for computing scanpath similarity – here we review a representative sample.

Early methods for measuring the similarity between two scanpaths calculated the absolute distance between the scanpaths' fixation coordinates (e.g., 'Mannan Linear Distance', (Mannan, Ruddock, and Wooding, 1995)). 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 according to which AOI they appeared in and scanpaths were represented as strings of AOI labels. 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 (Jarodzka 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: (1) shape is the difference in two saccade vectors u_i - v_j, (2) direction is the difference in angle between the saccade vectors, (3) length is the difference in amplitude of saccade vectors $\| u_i - v_j \|$, (4) position is the distance between fixations (using absolute coordinates: x, y), and (5) duration is the difference in duration between fixations. Recently, analyses comparing MultiMatch and Scanmatch have reported advantages for MultiMatch in accuracy for measuring similarity (Dewhurst et al., 2012; Foerster and Schneider, 2013; Gurtner, Bischof, and Mast, 2019). Thus, given that MultiMatch measures several scanpath features, it will be used it as the similarity tool for the present research. We now review a representative sample of related research on scanpath analaysis.

2.2 Scanpaths from Image Viewing

One of the early sources of scanpath data was from image viewing tasks. Traditionally, this work relied on saliency map models to inform predictions of eye movement patterns during image viewing. A saliency map is the representation of an image that encodes the importance of each area in the scene according to low level features. Saliency maps were used as inputs to computational models, which predicted where people would look in the image. These models embedded the hypothesis that shifts of attention would be directed towards salient locations in the image. However, when compared with data from human participants, the accuracy of these models was low (Jovancevic, Sullivan, and Hayhoe, 2006; Turano, Geruschat, and Baker, 2003; Stirk and Underwood, 2007).

Foulsham and Underwood (2008) sought to improve prediction of eye movements during scene and image analysis by evaluating a theory that a series of fixations and saccades, i.e., a viewing pattern, are stored in memory along with a representation of the image. Known as the scanpath theory, it posits that when an individual is shown a previously viewed image for a second time, they will engage in a similar viewing pattern and this will facilitate recognition of the image (Noton and Stark, 1971). Importantly, this theory proposes that saliency alone is not the key predictor where people will look in an image, and that individual viewing biases will be better predictors of scanpath similarity for an image viewed at two different time points. In the experiment, Foulsham and Underwood (2008) recorded eye movements while participants viewed scenes at two points in time: when the scene was first encountered (encoding phase) and subsequently when they saw the same scene again later in the experiment (recognition phase). Using Levenshtein's string edit distance (Brandt and Stark, 1997; Hacisalihzade, Stark, and Allen, 1992) and Mannan similarity measure (Mannan, Ruddock, and Wooding, 1995), each participant's scanpath at the encoding phase was compared against their scanpath at the recognition phase. Their scanpaths were also compared to simulated scanpaths for same image created by leading saliency model at the time (Itti and Koch, 2000). Results showed that although scanpaths from the encoding phase had high similarity to simulated scanpaths from saliency model, they were even more similar to the scanpath produced by viewing the same image at the recognition phase. This demonstrates that viewing patterns could not be predicted by image saliency alone and that provided evidence that incorporating scanpath theory could be used to improve vision model predictions.

In general, in order to have a more accurate method of predicting scanpaths, top-down mechanisms need to also be accounted for. These top-down mechanisms correspond to anything guiding attention that is not directed by the image features or low-level viewing mechanisms. Thus, these can include characteristics of the observer (e.g., age, culture, gender, health) as well as of the task (e.g., reading, image search, face viewing, problem solving). Above, we described that the bias of observer-related characteristics played a role in influencing viewing patterns (Foulsham and Underwood, 2008). More recently, Coutrot, Hsiao, and Chan (2018) tested a novel method for scanpath modelling and prediction that focused on computing visual patterns of top-down mechanisms (e.g., experimental task being completed) and low-level features (i.e., stimuli-related information, like what is being looked at). Their approach involved hidden Markov models (Chuk, Chan, and Hsiao, 2014) to model eye movement and Discriminant Analysis to classify the simulated scanpath. Coutrot, Hsiao, and Chan (2018) tested their approach with two datasets. The first dataset contained scanpaths from participants engaged in several activities: (1) free viewing, where participants examined an image without further instruction, (2) saliency search, where participants had to judge which half of an image was more salient, and (3) cued object search, where participants had to determine whether target object was present (Koehler et al., 2014). The second dataset contained scanpaths from scene viewing during which music was either present or not (Coutrot and Guyader, 2014). The Markov model was able to correctly identify the viewing task in the first data set and whether music was present in the second dataset. Thus, Coutrot, Hsiao, and Chan (2018) demonstrated that a single analysis tool could discriminate viewing patterns that arose from both observer and stimuli influences.

Other work focuses on investigating how saliency influences the visitation order of salient areas and so adds to our understanding of how people select for these regions. To illustrate, Wang et al. (2017) proposed a new scanpath estimation model that incorporates various dynamic saliency features, such as inhibition of return, where most recent fixation locations have lower probability of being revisited (Posner and Cohen, 1984; Posner et al., 1985). By taking these dynamic influences into consideration, a probability map of gaze shifts can be derived, which generates potential candidates for the next fixation in a given fixation sequence. Thus, the model predicts scanpaths, fixation by fixation. In order to evaluate the performance of their proposed scanpath prediction model, Wang et al. (2017) compared it against four existing saliency models, using two eye tracking data sets collected by asking participants to view various scenes (Bruce and Tsotsos, 2005; Judd et al., 2009). Each experimental scanpath was compared to the simulated scanpaths predicted by the different saliency models using several scanpath comparison methods (e.g., ScanMatch, described above, and Hausdorff distance, which computes the maximum value of all minimum distances between two sets of scanpaths (Wang et al., 2011)). The new model for scanpath estimation showed significantly higher similarity to actual participants' scanpaths in both datasets compared to previous models. The results show that scanpath estimation is furthering the understanding of dynamic saliency features and how they impact scene viewing.

2.3 Scanpaths in Different Populations

In addition to scene analysis, scanpaths have been used to distinguish populations for a variety of cognitive tasks. To illustrate, French, Glady, and Thibaut (2017) employed scanpath analysis to investigate whether children and adults used different strategies when solving analogy problems. Participants in both age groups were presented with images containing characters that conveyed a specific context (for instance, a cat chasing a mouse) and the participants had to view the image in order to identify and respond with the target "missing" relationship (for example: A is to B as C is to ?). A given participant's scanpath for each image was compared all other participants' scanpaths for that image and similarity was calculated using a variety of tools, including Levenshtein's String Edit algorithm (Levenshtein, 1966), the MultiMatch comparison tool (Jarodzka et al., 2012) and an Attentional Map (AMAP) comparison (Jost et al., 2004; Rajashekar et al., 2008). Scanpath similarity scores for each tool were placed into matrices, so each matrix consisted of child-child, child-adult and adult-adult scanpath similarity scores. These matrices represent high dimensional data that make analysis challenging. To address this, a multi-dimensional scaling procedure (Torgerson, 1952) was applied to the matrices to reduce dimensionality to a 2D space (a x, y scatterplot). In that graph, the location of a given scanpath represented how similar/dissimilar it is to all other scanpaths (e.g., if two scanpaths had a high similarity score, they would be located close together in the 2D scatterplot). This data was used as input to a neural network classifier that predicted whether the input data belonged to a child or an adult. The classifier performed above chance, with the MultiMatch comparison tool outperforming the other tools.

McIntyre and Foulsham (2018) were interested in cultural effects on visual attention, specifically while teaching. Previous research found that as teachers gain more experience, they spend more time looking at their students. However, much of the previous research has been conducted exclusively in Western settings. To investigate the impact of culture and teaching expertise on attentional patterns in the classroom, McIntyre and Foulsham (2018) recorded eye movements from both novice and expert teachers in the UK and in Hong Kong as they instructed a ten-minute unit. Eye movements were coded according to the teacher's didactic activity at the time, namely lecturing the students or facilitating discussion, and subsequently, Levenshtein's String Edit algorithm (Levenshtein, 1966) was used to calculate scanpath similarity. To examine effects of expertise, three types of analyses were conducted: (1) expertise, where scanpaths were compared within a given expertise group (expert – expert, novice – novice) and across expertise groups (expert – novice); (2) culture, where scanpaths were compared within a given culture group (UK – UK, Hong Kong – Hong Kong) and across culture groups (UK – Hong Kong); and (3) a further subdivision into four different categories of comparisons: same culture-different expertise, different culture-same expertise, different culture-different expertise, same culture-same expertise. Overall, the similarity between scanpaths converged towards a single viewing pattern as expertise increased, where expert teachers in both cultures tended to direct more focused attention towards individual students than novice teachers. Furthermore, a cultural difference was found, where teachers from Hong Kong employed more scanning of the student audience while lecturing compared to the UK teachers.

Other studies have used scanpath analysis to compare experts and novices in medical studies. Crowe, Gilchrist, and Kent (2018) extracted scanpaths produced during tumour diagnosis. The scanpaths came from three populations: undergrads with no experience, medical students with background knowledge, and practicing doctors with extensive expertise. ScanMatch was used to compare scanpaths within each level of experience. For each brain slice image, a participant's scanpath was compared to all other participants' scanpaths for that image. Similarity scores were grouped into within-condition groups (expert-expert, medical student-medical student and undergrad-undergrad) as well as across-condition groups (i.e., expert-undergrad or expert-medical student). For within-expertise level groups, experts had the most similar scanpaths, and thus achieved the highest consistency of viewing patterns during diagnosis. Interestingly, medical students had least similar scanpath patterns during diagnosis, even more so than undergraduate students with no experience. One explanation for this finding is that the undergraduates focused on the salient regions as they had no knowledge to guide them, and these salient regions led to consistent viewing patterns. In contrast, the medical students had some knowledge and that affected their visual attention, but because the knowledge had not yet solidified, their viewing was not consistent. These results emphasize the importance of examining how distribution of attention is affected at different stages of learning.

As yet another example, Castner et al. (2018) investigated how stages of expertise influenced visual attention of dentistry students. Scanpaths were acquired from students at different stages of learning to read dental x-rays. Using supervised learning classifier tools (SubsMatch and Needleman-Wunsch), they found that scanpaths from students at pre-training were distinctly different than expert scanpaths elicited from latter semester students. Six weeks later, at post-training, students' scanpaths were similar to expert scanpaths. However, they were also equally as similar to students' scanapths from middle semesters, which showed little increase in scanpath similarity. This indicates that exploratory behaviour undergoes considerable change at early stages of knowledge acquisition but this settles into more consistent patterns at latter stages of learning. Without scanpath analysis, it would have been difficult to fully illustrate the dynamic change in visual attention throughout development of expertise.

2.4 Scanpaths Distinguish Experimental Tasks

In addition to distinguishing between populations viewing various stimuli, scanpaths have been used to distinguish between different experimental conditions. Given that eye movements are affected by type of image viewed, scanpath patterns should reflect differences in experimental stimuli. Dewhurst et al. (2018) tested this prediction using MultiMatch tool and data from a visual search task.

During the viewing task, for a given trial, numbers 1 through 5 were presented in random locations on a single screen (this is referred to as an "image" below) and participants were instructed to fixate on each number in ascending order. While the locations were random for a given image, this image was shown to all participants. The goal was to control for individual variation in viewing patterns and create highly similar scanpaths as each trial should elicit five fixations in similar locations and same order. Task difficulty was manipulated across five levels by increasingly obscuring the viewing area, which rendered the target numbers difficult to find. Each scanpath for a given participant and trial was compared to the scanpaths of all other participants for that same image. Thus, this was a between-participant analysis but within task-analysis. Dewhurst et al. (2018) found that easy search tasks resulted in more similar scanpath similarity scores whereas search tasks that were obscured produced lower similarity scores. The authors concluded that as perceptual difficulty of a task increases, scanpaths become less similar. Thus, task difficulty reduces the consistency of viewing patterns.

Zhou et al. (2016) examined whether scanpath similarity analysis could distinguish cognitive processes during risky decision-making tasks. Prior research indicated that decision making was influenced by level of risk, so the authors aimed to determine if this finding was also reflected in viewing patterns captured by scanpaths. They used eye-tracking data collected from three different studies. Each study consisted of two within-subject conditions: baseline tasks for the control condition and risky tasks for experimental condition. The details of the control and experimental stimuli varied per study, but the targeted cognitive processes associated with risky decision-making were similar. The analysis of scanpath similarity scores was set up to investigate whether scanpath similarity within each condition was higher than between the two conditions. This was done by comparing all possible pairs of scanpaths using ScanMatch (Cristino et al., 2010) and organizing the similarity scores into three groups: similarity scores that resulted from comparing two scanpaths from the control condition were placed into the within-control group, similarity scores that resulted from comparing two scanpaths from the experimental condition were placed into the within-experimental group. Finally, each score that resulted from comparing scanpaths between the two conditions was placed in the across-condition group.

In the present thesis, we rely on the analysis approach outlined above, consisting of computing within and across condition similarity scores, as we describe in further detail in Chapters 3 and 4. Zhou et al. (2016) used multi level modeling with within/across group as the fixed factor and participant as the random factor. Results showed that (1) the scanpath similarity scores were high in both within-control and within-experimental groups, indicating that both types of decision tasks elicit consistent viewing strategies, and (2) the mean similarity scores are lower in the across-condition group compared to the mean within-condition scores, demonstrating that the different risk levels involved in decision-making tasks also elicited different viewing strategies.

2.5 Scanpaths for Characterizing Viewing Patterns

The studies described thus far have focused on analyzing the similarity of scanpaths coming from different populations and/or experimental conditions. Other studies have instead utilized scanpath data to characterize viewing behaviors in order to shed light on cognitive processes that might otherwise be difficult to capture. 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 pattern requires more than fixation data alone, the analysis involved data on sequences of fixations, i.e., scanpaths. In order to measure presence of pre-defined viewing patterns, Holmqvist et al. (2011) devised a novel algorithm. Rather than measuring similarity between two scanpaths, the algorithm characterized eye movements' exploration of the areas of interest (AOI) in terms of overview or focused viewing patterns, and measured how the scanning behaviour changed over time. Just as in the string edit method, a sequence of letters was created to represent change in AOI's fixated upon per scanpath. Subsequently, a 'sliding window' mechanism was used that counted how many different areas where looked at within a smaller sequence of fixations. Holmqvist et al. (2011) found that high-ability students had significantly more focused scanning patterns as compared to low-ability students.

Gurtner, Bischof, and Mast (2019) analyzed scanpaths to determine if there were differences in visual attention between viewing scenes in real-life versus recalling the same image in memory. The method used was Recurrence Quantification Analysis (RQA), which measures return of the eye gaze to previous fixations. The results showed that when mentally recalling an image, our eye movements tend to re-fixate earlier and more often compared to actual scene viewing. Zhang, Anderson, and Miller (2021) also used the RQA measures to examine the effect of mind wandering on visual attention while examining different imagery. The results showed that during mind wandering participants' revisit earlier fixations as if to review stimuli, resulting in duplicated scanpaths. Without measuring temporal dynamics of attention these effects of mental imagery on gaze patterns could not have been detected.

2.6 Summary

To date, scanpath analysis work has focused on analyzing whether visual gaze patterns can distinguish between tasks or populations. Indeed, scanpath analysis has increased the accuracy of predicting what is attended to during scene viewings and when they are attended to. Additionally, scanpaths have made it possible to discern viewing patterns particular to observer-related and task-related information and discriminate between experimental conditions. However, scanpath analysis is a young field and there is still relatively little research examining its full potential. This is especially the case in educational settings that investigate the impact of instructional manipulations on student performance and learning. We take a preliminary step in addressing this gap, by exploring the utility of scanpaths for distinguishing experimental conditions in eye-tracking data from two studies in the math domain: (1) Study 1 manipulated problem format in a basic arithmetic task, while (2) Study 2 manipulated example format in a more complex algebra problem-solving task.

3 Scanpath Analysis 1, Basic Division Problems

3.1 Introduction and Background

In this Chapter, we present scanpath analysis 1 [1]. The analysis used data from a previous study (Tan, Muldner, and LeFevre, 2016) that examined how students mentally solved basic division problems, such as 30 $\div$ 5 = [ ]. Thus, we begin with some background to provide context.

While there are various strategies that students can use for solving basic division problems, early work proposed that once a student had sufficient practice, problems were solved simply by recalling the answer directly from memory (e.g., Groen and Parkman, 1972; Siegler, 1989). This solving strategy was called 'direct retrieval'. In order for direct retrieval to be possible, arithmetic knowledge must be stored in memory as 'facts' such as $2 + 4 = 6$. This fact is composed of two operands (2 and 4), one operator (+), and a solution (6). An arithmetic problem is an incomplete fact, where (most often) the solution is missing. To solve the problem through direct retrieval, the student must search their memory for the arithmetic fact that matches the given problem and retrieve the solution from this mental representation of the fact (Ashcraft, 1992; Campbell and Oliphant, 1992; Widaman et al., 1989; Widaman et al., 1992).

[1] A subset of this work was presented in Stranc, Tan, and Muldner (2020)

Direct retrieval was hypothesized to be the key method for solving basic arithmetic problems (Groen and Parkman, 1972; Siegler, 1989), that is until the problem size effect was identified. The problem size effect describes the phenomenon where arithmetic problems with operators bigger than 10 (referred to henceforth as 'large problems') take longer to solve compared to problems with operators smaller than 10 (referred to henceforth as 'small problems'). This finding was unexpected because if problems were solved through retrieval, the size of the operators should not have an impact. LeFevre et al. (1996) proposed that the difference in solving time was due to various strategies being used to solve the arithmetic problems. To tests this, students were asked to self-report on how they solved both large and small arithmetic problems. The results demonstrated that various solution strategies were indeed being employed. For instance, students reported breaking up some of the operands (e.g., 16 + 4 =_ became 10 + 6 + 4 =_) or using tricks and short cuts (where 3 × 9 = _ became (3 × 10) – 3 = _). Importantly, when students reported using direct retrieval for either small or large problems, they were faster compared to other 'non-direct' retrieval solving strategies. Small problems (and well-practiced problems) are more likely to be solved by direct retrieval (Mauro, LeFevre, and Morris, 2003). Perhaps because large problems are less practiced, they are more likely to be solved through strategies other than direct retrieval.

Here, we focus on division problems, as they are most relevant to the present work. Mauro, LeFevre, and Morris (2003) suggested that one strategy for solving division problems involved casting them into a multiplication format and then directly retrieving the answer for the recast problem from memory. This two-step strategy was called "mediated retrieval" because the solution to a division problem was mediated by the solution for a multiplication problem. Mediated retrieval is most likely to occur when the solution is not directly retrievable.

Compared to direct retrieval, mediated retrieval generally takes longer and has lower accuracy due to the extra step in processing (Mauro, LeFevre, and Morris, 2003). The type of strategy used to solve a problem also impacts visual attention. During direct retrieval, across various types of arithmetic problems, students tend to pay more attention to the operator (+, -, ×, ÷) (Huebner and LeFevre, 2018). In contrast, solving strategies that involve mental transformations result in increased attention to the operands (Huebner and LeFevre, 2018). This latter finding may be the result of students' mental manipulation of these operands during the solving process. However, to date not many studies have employed eye tracking to investigate this phenomenon.

In summary, when a student solves a simple division problem and cannot directly retrieve the solution from memory, they use strategies that take more time – a common strategy is mediated retrieval. Recently, Tan, Muldner, and LeFevre (2016) investigated the nature of the two-step process involved in mediated retrieval for small and large division problems. They presented students with division problems in *traditional* format, with the division operator present (21 ÷ 7 = _) as well as division problems recast into their equivalent multiplication format, with the multiply operator present (7 × _ = 21). Participants' visual attention was captured by an eye tracker and the subsequent data analysis focused on dwell time and fixations. The goal of the present thesis research is to extend the Tan, Muldner, and LeFevre (2016) results with analysis of participants scanpaths using the MultiMatch tool. We begin by briefly describing the Tan, Muldner, and LeFevre (2016) study.

3.2 Overview of Tan, Muldner, and LeFevre (2016) Study

3.2.1 Design, Methods and Procedure

In the Tan, Muldner, and LeFevre (2016) study, students solved basic division problems as an eye tracker captured their attention in two different presentation formats: *traditional* format (Table 3.1, rows a + b), and *recast* format, which used the multiplication format (Table 3.1, rows c + d). The study used a within-subject design to present each participant (*n* = 33) with 144 simple division problems (72 in *traditional* format and 72 in *recast* format). The problem order was randomly shuffled for each participant. Two factors were manipulated for each problem format. The first factor corresponded to the position of the missing element (either in the 3rd or the 5th position in the equation for each format type, see Table 3.1; note that although the missing element varied, the dividend was always located in the first position). The second factor corresponded to the operand size, with small problems containing dividends smaller than 25 and large problems containing dividends equal to or greater than 25. An eye tracker captured visual attention as participants solved the division problems. Participants were asked to state their response verbally for each problem and time taken to solve each problem was measured. The experiment lasted approximately 50 minutes. For the analysis, five areas of interest (AOIs) were drawn around each of the operator and operand items as shown in Figure 3.1.

3.2.2 Key Results from Tan, Muldner, and LeFevre (2016)

Overall, division problems presented in *recast* format were solved more quickly than those in *traditional* format. In both presentation formats visual attention

Table 3.1: Division Problem Formats

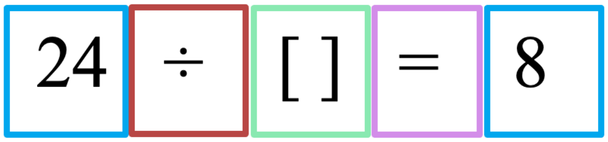
Traditional division format					
(a)	72	÷	[ ]	=	9
(b)	72	÷	9	=	[ ]
Recast multiplication format					
(c)	72	=	[ ]	*	9
(d)	72	=	9	*	[ ]

Note: The four different problem formats organized by format type (*traditional* and *recast*) as well as position of missing element (3rd or 5th position)

was focused on the left and middle problem regions, or the first three AOIs (see Figure 3.2). This is not surprising as the latter part of the problem equation did not offer pertinent information regarding the arithmetic fact. Problem format did, however, impact the distribution of attention for the first three AOIs. When participants solved problems in the *recast* format, the middle operand (third AOI) earned longest total dwell time compared to the other problem elements (see Figure 3.2). In contrast, when participants solved problems in the *traditional* format, fixations were more evenly distributed across the first three problem elements, as illustrated in Figure 3.2. While the eye tracking analysis focused on dwell and fixation time, some qualitative analysis was done to examine attention patterns over time. This qualitative analysis showed that there were differences in viewing patterns between formats but a scanpath analysis was not conducted.

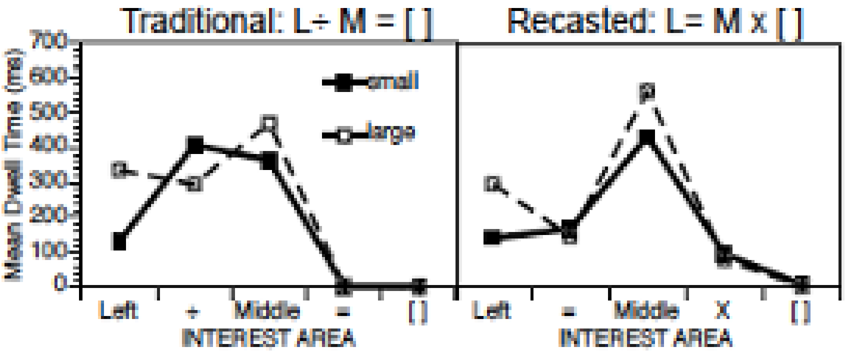
In summary, Tan, Muldner, and LeFevre (2016) found a difference in gaze distribution between the *traditional* and *recast* problem formats and hypothesized it was due to the use of different strategies. Importantly, the eye fixation analysis provided some evidence that students were using two different solving strategies depending on problem for the two different problem formats.

Figure 3.1: *Example of a Division Problem in Traditional Format*



Note: Five different areas of interest (AOIs) are outlined here around each operator and operand, including measurements. Note that the AOI labels were not visible to participants.

Figure 3.2: *Fixation Duration Results from Tan, Muldner, and LeFevre (2016)*



Note: Mean total dwell time during the solving process for all five AOI's in both *traditional* and *recast* problem formats.

3.3 Present Scanpath Analysis

The study by Tan, Muldner, and LeFevre (2016) included basic fixation and dwell time analysis to capture visual attention. The goal of the current analysis was to extend this analysis to examine how scanpaths were affected by problem presentation format that was manipulated in the original study. We used the Multimatch Tool for scanpath analysis.

Our research question was as follows: Does division problem format impact scanpath similarity? We anticipate that the two problem formats will affect the distribution of visual attention (and thus scanpath similarity). If we can demonstrate that scanpaths are indeed different between the two conditions, this provides indirect support that the conditions elicit different solving strategies.

3.3.1 Data Preparation and Analysis Framework

Data preparation. In preparation for data analysis, we extracted the scanpaths from the eye tracker file, as follows. Recall that each participant solved 72 *traditional* and 72 *recast* division problems while an eye tracker logged their eye fixations. We extracted every participants' scanpath for each problem they solved. A given scanpath corresponded to the series of fixations produced while the problem was solved and was stored in a single text file which was labelled with the participant number and problem ID. Thus, each file contained a single scanpath and was composed of a series of rows corresponding to the number of fixations belonging to the scanpath (one fixation per row); each row included the x and y coordinates for the fixation, as well as the duration of that fixation (in seconds). Only fixations that fell within the five AOIs (see Figure 3.1) were included in a scanpath file. Due to the requirements of MultiMatch tool, scanpaths that had fewer than three fixations were not included in the analysis. 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). We included both fifth and third missing element positions (see Table 3.1) in the analysis.

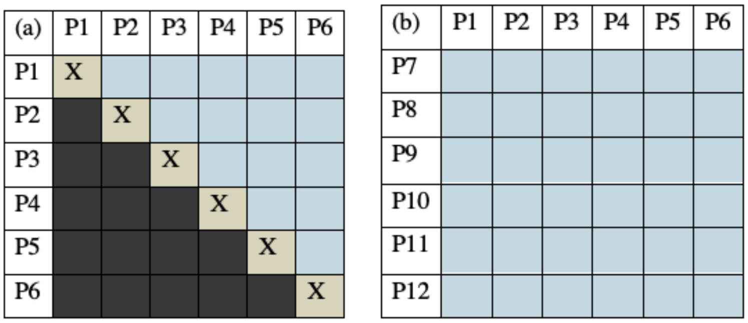
The MultiMatch tool we used for scanpath analysis includes the option to group fixations within the scanpath, which simplifies the analysis. However, due to the short scanpath lengths in the current data, simplification risks eliminating relevant fixation data. Thus, we turned off simplification thresholds that MultiMatch requires the dimension of the computer screen, which was set at a resolution of 1024 x 768 based on the computer screen used in the original study.

Analysis framework. Recall that each participant was exposed to two conditions that corresponded to the two problem formats: *traditional* and *recast*. Once the scanpaths were extracted, we followed the methodology from previous work (Dewhurst et al., 2018; Zhou et al., 2016) to compare scanpaths using MultiMatch. This approach involves comparing pairs of scanpaths within each condition to each other and also comparing scanpaths across 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 across-condition comparisons (Mathôt et al., 2012).

To obtain an overall measure of scanpath similarity in a given condition, we wrote a Python script that called the MultiMatch library to calculate scanpath similarity scores. For the within-condition comparison, for every participant, the script compared each scanpaths in a given condition to each other scanpath in that condition - each comparison corresponded to a pair of scanpaths. In other words, the script compared all of a participant's scanpaths in the *traditional* format; followed by a second part of the script that compared all the participant's scanpaths in the *recast* format. Note that order of comparison does not matter, i.e., similarity (scanpath1, scanpath2) = similarity (scanpath2, scanpath1). Thus, the within 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 as shown in graph (a) in Figure 3.3 (e.g., if there were only 4 scanpaths for a given format, this would produce 4!/4!(n-2)! Similarity scores). For every pair of scanpaths that was compared, MultiMatch produced five similarity scores ranging from 0 (no similarity) to 1 (perfect similarity), one for each feature (shape, direction, length, position and duration). The script then computed the mean similarity score for each participant for a given condition (*traditional* or *recast*) by summing all the similarity scores in that condition for that participant and dividing it by the number of scores. For each of the five features, this resulted in a total of 33 mean similarity scores for the *traditional* within comparison and 33 mean similarity scores for the *recast* within comparison (as there were 33 participants and each solved problems in both formats).

The script also computed an across-condition comparison. This comparison works similarly to the within-condition comparison described above, except that now pairs of scanpaths were compared from different problem formats. Specifically, for a given participant, each scanpath from every problem this participant solved in the *traditional* format was compared to every scanpath from problems they solved in the *recast* format. This across-condition comparison involves n1 x n2 comparison scores, where n1 and n2 correspond to the number of problems in each condition (see graph (b) in Figure 3.3). The script computed the across-condition similarity score for each participant and each of the five MultiMatch features resulting 33 x 5 across-condition comparison scores.

Figure 3.3: *Comparison Matrices*



(a)	P1	P2	P3	P4	P5	P6
P1	X					
P2	■	X				
P3	■	■	X			
P4	■	■	■	X		
P5	■	■	■	■	X	
P6	■	■	■	■	■	X

(b)	P1	P2	P3	P4	P5	P6
P7						
P8						
P9						
P10						
P11						
P12						

Note: Graph (a) illustrates the correlational matrix for a within condition analysis, comparing scanpaths within a condition. Graph (b) illustrates the comparison matrix for an across condition analysis, comparing scanpaths between two conditions.

In sum, the comparison analysis produced three groups of MultiMatch similarity scores for each feature, which we refer to as Similarity$_{\text{traditional}}$, Similiarity$_{\text{recast}}$ and Similarity $_{\text{across}}$.

The similarity scores were analyzed using inferential statistics. Current guidance on interpreting the results from this type of analysis framework is limited. One guideline is that if different cognitive processes are used for the two problem formats, then the two within-condition similarity scores, Similarity$_{\text{traditional}}$ and Similiarity$_{\text{recast}}$, will be lower than the Similarity$_{\text{across}}$ (Zhou et al., 2016; Mathôt et al., 2012). Moreover, if the two within-condition scores are significantly different, this implies that the conditions elicit different patterns of visual attention (i.e., scanpaths). In Chapter 5, we conduct simulations to provide further guidance on interpretation scanpath results. Prior work has also suggested 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$_{\text{traditional}}$ scores to on average be lower than the Similarity$_{\text{recast}}$ scores. On the other hand, the *recast* format is less common and thus arguably more challenging – if that is the case, we could expect the Similarity$_{\text{recast}}$ scores to on average be lower than within the Similarity$_{\text{traditional}}$ scores.

In general, we leave detailed interpretation of the findings until the final discussion, because we will incorporate the results of the simulations we conducted that will help explain the results.

3.3.2 Results

Unless otherwise stated, the results are based on data from the 33 participants in the original study. Recall that we had three collections of scanpath similarity scores, Similarity$_{\text{traditional}}$, Similarity$_{\text{recast}}$ and Similarity$_{\text{across}}$. As noted previously, these scores were generated by MultiMatch, which produces a similarity score (between 0 and 1) for five features per scanpath comparison. The descriptives for each feature are in Table 3.2.

We followed up the descriptives with inferential statistics. For each of the five MultiMatch features, we conducted a separate one-way ANOVA with comparison type as the three level within-subject factor (Similarity$_{\text{traditional}}$, Similarity$_{\text{recast}}$, Similarity$_{\text{across}}$) and similarity scores for the target feature as the dependent variable. Thus, in this analysis for each participant there were three similarity scores (i.e., Similarity$_{\text{traditional}}$, Similarity$_{\text{recast}}$, Similarity$_{\text{across}}$). Sphericity violations were corrected using the Greenhouse-Geisser adjustment. Significant effects were followed up with pairwise comparisons with Bonferroni correction.[2]

[2] The confidence intervals for the graphs in Figures 3.4, 3.5, 3.6, 3.7 and 3.8 were calculated using the method advocated in Field (2013) for repeated measure designs.

Table 3.2: Descriptives for Analysis 1 Similarity Scores

	Shape		Direction		Length		Position		Duration	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ALL										
Recast	0.9387	0.015	0.7707	0.087	0.9226	0.015	0.8965	0.017	0.637	0.063
Traditional	0.9448	0.016	0.7687	0.101	0.9296	0.018	0.9004	0.021	0.6256	0.057
Across	0.9391	0.014	0.7599	0.095	0.9222	0.015	0.8918	0.015	0.6287	0.057
POSITION_3										
Recast	0.9425	0.013	0.7685	0.104	0.9244	0.016	0.8976	0.02	0.6354	0.062
Traditional	0.9381	0.021	0.766	0.111	0.9224	0.019	0.8913	0.024	0.637	0.065
Across	0.9401	0.016	0.7652	0.095	0.9235	0.017	0.894	0.019	0.633	0.061
POSITION_5										
Recast	0.9432	0.015	0.7744	0.089	0.9274	0.017	0.8997	0.019	0.6294	0.061
Traditional	0.9386	0.014	0.7674	0.106	0.9222	0.017	0.8942	0.018	0.6225	0.062
Across	0.9407	0.014	0.7716	0.084	0.9248	0.015	0.896	0.017	0.6262	0.061
FAST										
Recast	0.9045	0.024	0.6487	0.101	0.9217	0.018	0.853	0.031	0.7186	0.07
Traditional	0.9129	0.028	0.6655	0.132	0.9294	0.02	0.8499	0.063	0.7129	0.061
Across	0.9063	0.026	0.6536	0.091	0.9249	0.016	0.8469	0.043	0.7179	0.048
SLOW										
Recast	0.9265	0.031	0.7492	0.128	0.9227	0.015	0.8748	0.026	0.6517	0.055
Traditional	0.9329	0.032	0.7513	0.122	0.9292	0.018	0.8814	0.032	0.657	0.058
Across	0.9296	0.022	0.7517	0.1	0.9225	0.015	0.8711	0.023	0.6525	0.052

Note: ‘Recast’ rows represents Similarity$_{\text{recast}}$ scores, ‘Traditional’ rows represents Similarity$_{\text{traditional}}$ scores and ’Across’ rows represent Similarity$_{\text{across}}$ scores. The main analysis corresponds to rows labelled "ALL" and Position 3 and 5 indicate blank positions.

There was a significant main effect of comparison type on scanpath similarity for three MultiMatch features (shape: $F(2, 64) = 13.28$, $p < 0.01$; length: $F(2, 64) = 11.29$, $p < 0.01$; position:$F(2, 64) = 7.53$, $p < 0.01$) – see Figures 3.4, 3.6 & 3.7. For shape and length, the pattern of results was the same and follow up comparisons showed that for the within-condition comparison, $\text{Similarity}_{\text{traditional}}$ had significantly higher scanpath similarity scores than $\text{Similarity}_{\text{recast}}$ ($p < 0.01$) (see Figures 3.4 and 3.6). Thus, the *traditional* division problems produced more consistent scanpath patterns compared to division problems that were recast into multiplication format. For the across-condition comparisons, $\text{Similarity}_{\text{traditional}}$ had significantly higher scanpath similarity than $\text{Similarity}_{\text{across}}$ ($p < 0.01$) but the difference between $\text{Similarity}_{\text{recast}}$ and $\text{Similarity}_{\text{across}}$ was not significant. Interpreting this results is more challenging as currently there is no guidance on it – we thus delay doing so until the final discussion.

In contrast, for the position feature, the within-condition comparison scores were not significantly different. Both within-condition comparison groups had significantly higher scores than the across-condition comparison scores, $\text{Similarity}_{\text{traditional}}$ ($p < 0.01$) and $\text{Similarity}_{\text{recast}}$ ($p < 0.01$) (see Figure 3.7). The other main effects were not significant (direction: $F(2, 64) = 0.74$, $p = 0.462$; duration: $F(2, 64) = 3.63$, $p = 0.061$).

Follow Up Analyses. We conducted several follow up analyses. The *traditional* and *recast* problems in the original study varied the missing element location in the equation, i.e., “blank”. which was the placeholder for the solution (see [ ] in Table 3.1). 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_3, position_5 corresponding to the position of the blank) and re-ran the inferential statistics. The mean descriptives can be found in Table 3.2.

Figure 3.4: *Mean Similarity Scores for Shape Feature (error bars represent 95% confidence intervals)*

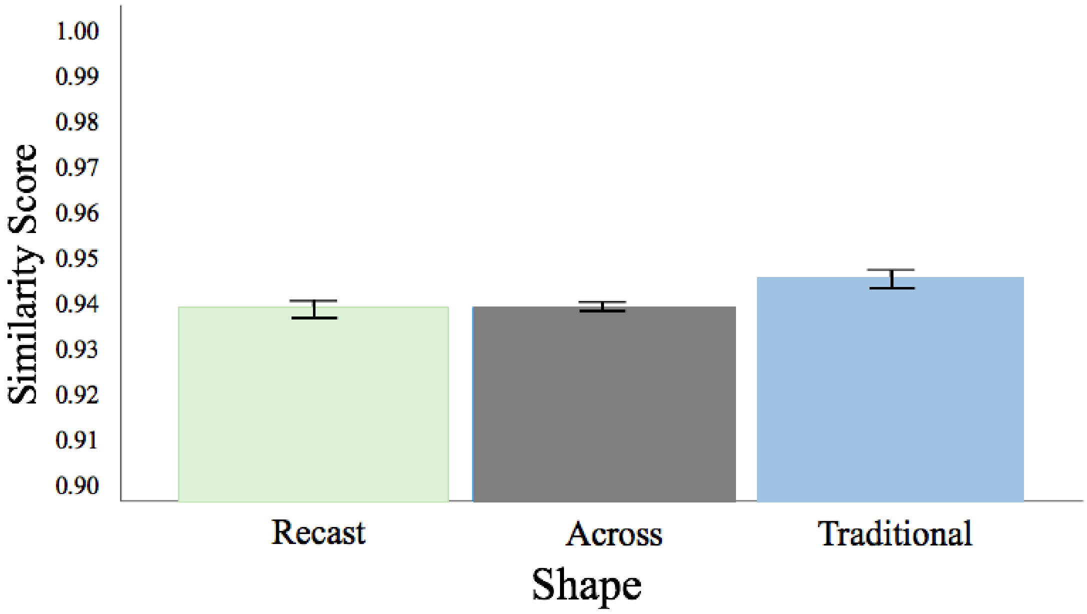
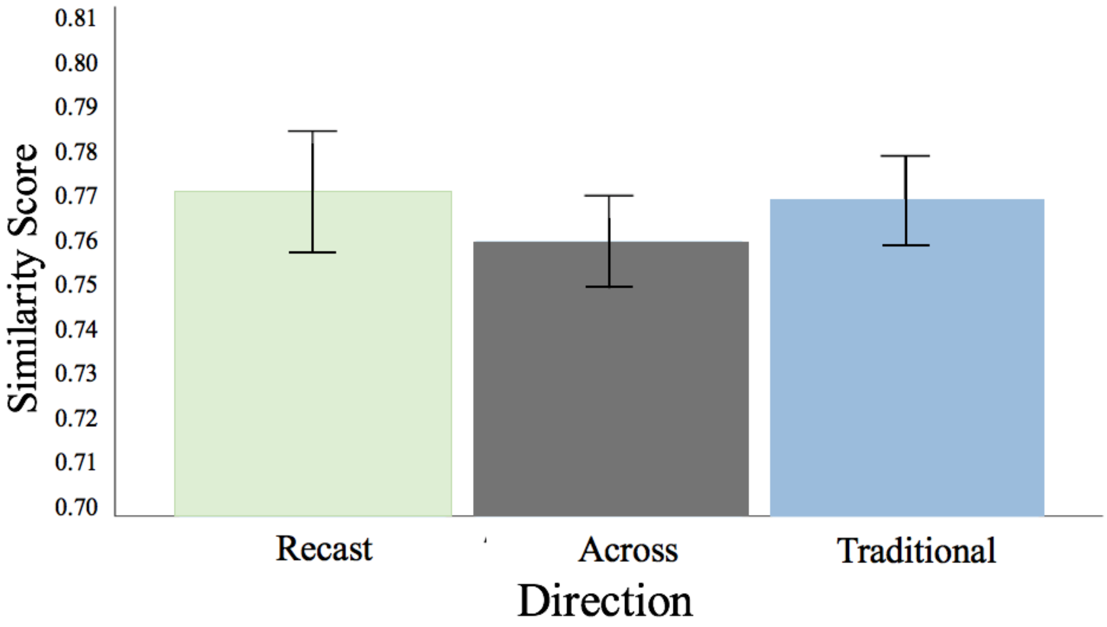


Figure 3.5: *Mean Similarity Scores for Direction Feature (error bars represent 95% confidence intervals)*



The main effect of blank position was not significant for any of the five features ($p > .139$), nor were there any significant interactions between blank placement and comparison group. The effect of condition, however, remained significant and held the same pattern as above (effect of shape, length, and position features resulted in significant main effects, $p < 0.05$).

We also 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 *recast* format). If participants solved problems primarily using retrieval, then there would be no differences in scores between $\text{Similarity}_{\text{traditional}}$ and $\text{Similarity}_{\text{recast}}$ groups, because retrieval does not require scanning and/or shifting visual attention between problem elements.

Figure 3.6: *Mean Similarity Scores for Length Feature (error bars represent 95% confidence intervals)*

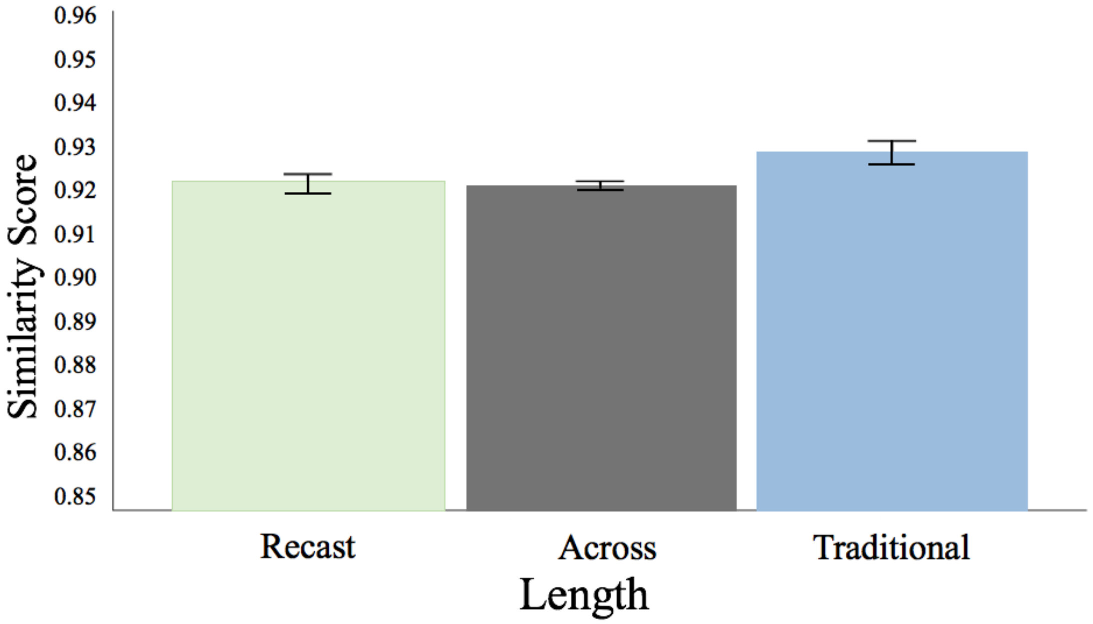
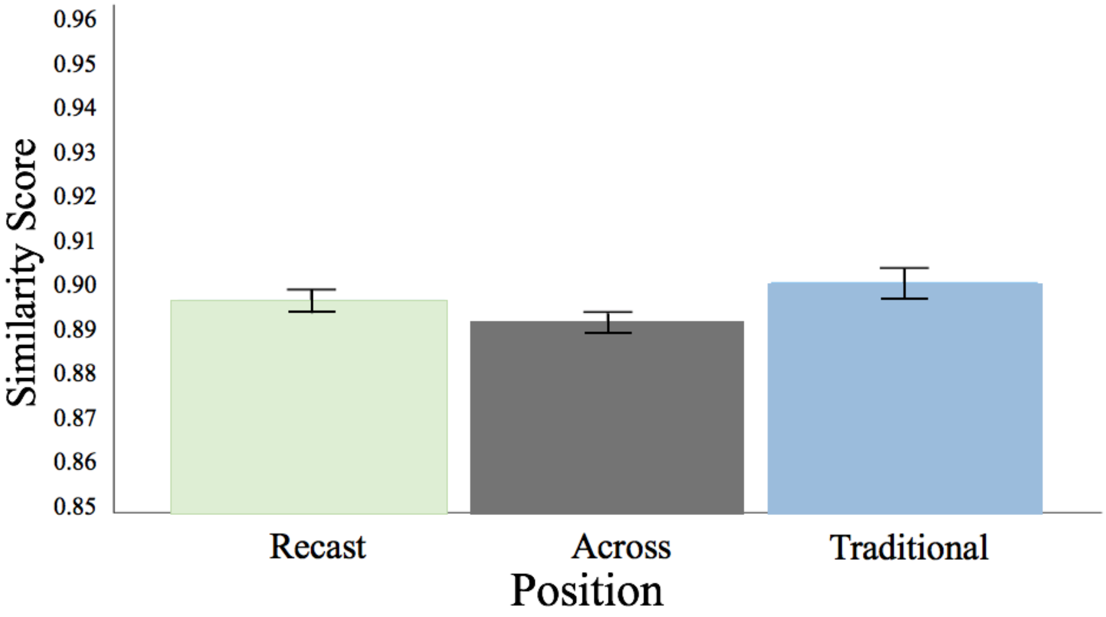


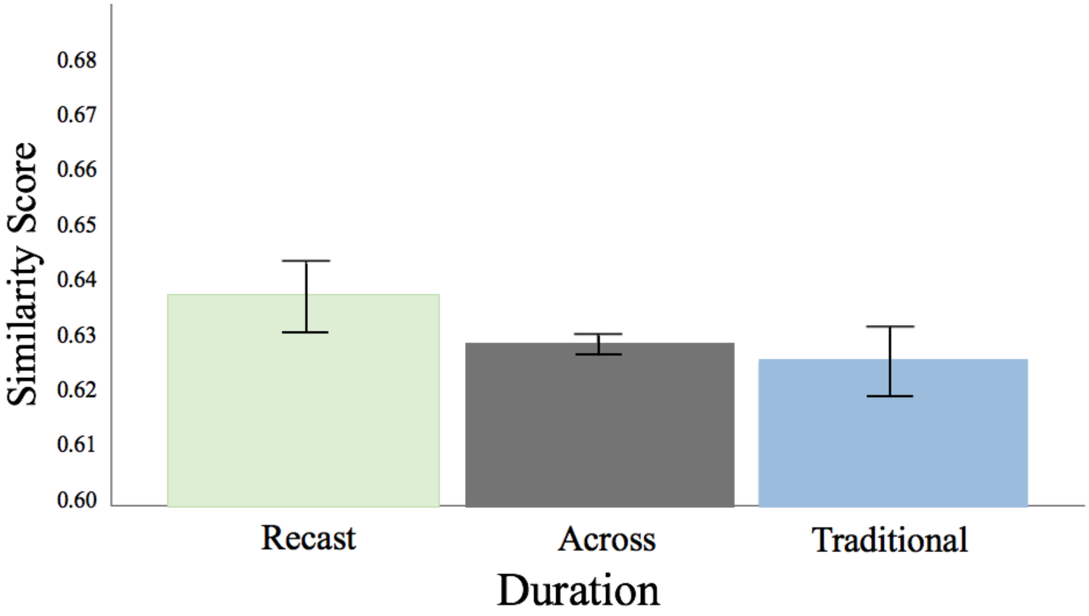
Figure 3.7: *Mean Similarity Scores for Position Feature (error bars represent 95% confidence intervals)*



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 ($\text{Similarity}_{\text{traditional}}$, $\text{Similarity}_{\text{recast}}$, $\text{Similarity}_{\text{across}}$) the mean score for slow scanpaths and the mean score for fast scanpaths. The descriptive mean results can be found in Table 3.2.

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

Figure 3.8: *Mean Similarity Scores for Duration Feature (error bars represent 95% confidence intervals)*



shape, length, and position, the being caveat that some of the follow up comparisons were no longer significant.

3.3.3 General Summary

In summary, the within and across condition analysis demonstrated that there were differences in eye movement patterns between the *traditional* and *recast* conditions. Features with significant differences all showed higher similarity scores for *traditional* compared to *recast* format. Two of these features (length and position) showed lower across-condition scores compared to within condition scores, which should indicate that the two conditions elicit different visual strategies. We will provide further interpretation guidelines in Chapter 5 and discuss the findings in light of these in Chapter 6.

4 Scanpath Analysis 2, Algebra Problems

4.1 Introduction and Background

In this Chapter, we present scanpath analysis 2[1] . The analysis used data from a study that presented algebra problems for students to solve that were paired with examples illustrating solutions to similar problems (Jennings and Muldner, 2020). Examples are commonly used in instructional materials such as textbooks, worksheets and online platforms. Learning from examples is more beneficial for learning than general procedure alone (Reed and Bolstad, 1991) or solving problems without examples (Sweller and Cooper, 1985). Examples are helpful because they illustrate not only the final solution but its step-by-step derivation, which students can use to help overcome impasses during problem solving. However, the effectiveness of examples depends on how students use them (Borracci et al., 2020; Muldner and Conati, 2010; VanLehn, 1998; VanLehn, 1999). There are two main solving strategies when examples and problems are both available, described below.

The first strategy students can use is to generate the problem solution by copying from the example. Students can copy the entire solution from the example or transfer only parts of it. Sometimes the example solution is transferred

[1] A subset of the results were presented in Stranc and Muldner (2020)

directly to the problem without any changes but often the student makes superficial changes, such as replacing example constants to match the constants required for the problem. Copying is more likely to occur when differences between the problem and example are easy to reconcile (Muldner and Conati, 2010; Reed, 2012; Reed, Dempster, and Ettinger, 1985). While copying does allow the student to quickly generate a solution to the problem, whether the solution is correct depends on the similarity between the problem and example. Moreover, as copying does not require information to be actively processed, copying interferes with students' ability to learn relevant domain rules. Without a deeper learning of the underlying concepts, students cannot apply the domain knowledge to a different problem without referring to another example.

Copying is a strategy that should be reflected in eye tracking data, albeit to date this has been shown in domains outside of problem solving. Studies investigating eye movements involved in copying of material from a source to the target indicate that regular shifts of attention occur between the source and target (Ballard, Hayhoe, and Pelz, 1995; Bulling et al., 2010; Bosse et al., 2014; Haselen, Van Der Steen, and Frens, 2000). According to a seminal study by Ballard, Hayhoe, and Pelz (1995), copying from a model (i.e., example) to the reconstruction (i.e., problem) area requires the storage of too much information in working memory to reproduce the entire model with one glance. Thus, multiple inspections of the model are required to complete the transfer of information, with gaze shifts occurring back and forth as long as it takes to reproduce the model. Bulling et al. (2010) also reported regular shifts of attention between the model and the reconstruction area during copying. In the context of learning from a paired problem and example, one could speculate that during copying repeated gaze shifts would occur between the problem and the example.

A second strategy for solving a problem presented alongside an example is to generate the problem solution by applying existing knowledge to generate the answer (VanLehn, 1998; Tricot and Sweller, 2014). There are two advantages of solving problems without copying: it helps strengthen knowledge through practice Muldner and Conati (2010) and it helps uncover knowledge gaps. Knowledge gaps lead to impasses, because problem solving cannot continue until the gap is resolved. Students have several options as to how to overcome an impasse.

One way to overcome an impasse is by relying on common-sense or overly-general reasoning to infer a new rule that addresses the impasse, either in the context of the problem or by studying the example (VanLehn, 1998; VanLehn, 1999). In the latter case, the worked-out example provides more information to the student compared to the former option (i.e., learning a new rule in the context of the problem without the help of an example). Once the student learns the new rule, it can subsequently be applied to similar problems without having to refer to an example for support.

As we did for the copying strategy, we now hypothesize on expected eye movements when a student is solving a problem without copying. Since there are no eye tracking studies related to this context, the following description of predicted eye movements is purely speculative. If the student already knows how to solve the problem, visual attention will focus on the problem area. However, if the student encounters an impasse, their gaze may remain within the problem area as they try to infer a new rule using overly-general and common-sense reasoning. Alternatively, the student's gaze could shift to the example if they choose to infer a domain rule from its solution. Overall, however, when students solve problems without copying, we expect more visual attention to remain within the problem area rather than switching between the problem and example areas.

In summary, when a student attempts to solve the problem instead of by copying, they engage in deeper processing that is beneficial for learning. Thus, researchers have investigated ways to discourage copying. One approach is to manipulate problem-example similarity. When a problem and example pair are similar to each other, this promotes copying, as compared to when the problem-example similarity is reduced (Borracci et al., 2020; Lee, Betts, and Anderson, 2015; Muldner and Conati, 2010). More recently, Jennings and Muldner (2020) investigated the impact of varying problem-example similarity over time. They characterised problem-example similarity as the level of assistance: high similarity corresponded to high assistance, because the example solution could be copied, while reduced similarity corresponded to reduced assistance, because rather than copying, students had to solve the problem without copying. The goal of the present analysis is to extend the Jennings and Muldner (2020) study to include an analysis of scanpath similarity using the MultiMatch tool. We begin by briefly describing the Jennings and Muldner (2020) study.

4.2 Overview of Jennings and Muldner (2020) Study

4.2.1 Design, Methods and Procedure

In this study, students used a basic educational technology to solve algebra problems. Each algebra problem was paired with an example designed to provide assistance during problem solving. Students were presented with 12 different problem-example pairs. Half of the problem-example pairs offered high assistance, where the example was highly similar to the corresponding problem and so its solution could be copied to generate the correct problem solution. The other half of the problem-example pairs corresponded to reduced assistance, where the example was less similar to corresponding problem, thus blocking the generation of a correct solution through copying. However, as the example solution required the same set of rules as the problem solution, it still provided support to the solver. Thus, in this study, assistance provided by the example for solving the corresponding problem was operationalized through the similarity between the problem and example (high similarity = high assistance; reduced similarity = reduced assistance).

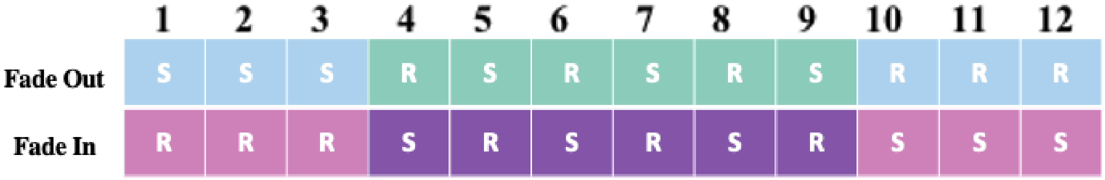
The study used a between-subject design with three conditions. We focus on the two conditions in which the timing of high vs. reduced assistance was manipulated:

- *Fade out* assistance (n = 20): participants were initially given high assistance problem-example pairs, but these transitioned to reduced assistance pairs after some problems were solved. As the assistance gradually moved from high to reduced, in this condition assistance faded out over time (see Figure 4.1, top).

- *Fade in* assistance (n = 19): participants were presented with reduced assistance problem-example pairs initially, eventually transitioning to high assistance pairs. Thus, in this condition assistance faded in rather than faded out (see Figure 4.1, bottom).

As shown in Figure 4.1, while the two conditions had the same number of reduced and high assistance problem-example pairs but the timing of assistance was varied. Participants were undergrad students who had not taken mathematics course(s) in university. An SR Research Eye Link 1000 eye tracker was used to capture gaze and fixations as participants solved problems and referred to examples. A basic software application was created to present the problems and examples shown in Figure 4.2. For each of the 12 problem-example pairs,

Figure 4.1: *Sequence of Presentation for 12 Problem-Example Pairs*



	1	2	3	4	5	6	7	8	9	10	11	12
Fade Out	S	S	S	R	S	R	S	R	S	R	R	R
Fade In	R	R	R	S	R	S	R	S	R	S	S	S

Note. Type of similarity for each of the 12 problem-example pairs in the *fade out* and *fade in* conditions; S represents high assistance pairs and R represents reduced assistance pairs.

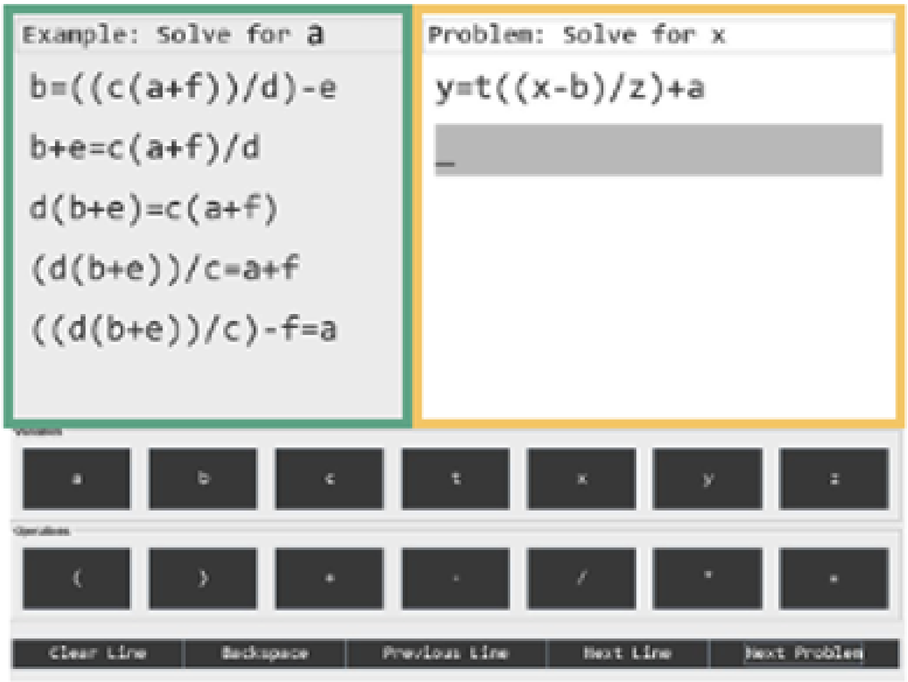
the application's interface presented the problem to solve on the right and a corresponding example on the left. In order to minimize head movement, the interface also included a virtual keyboard, which participants used to enter in solutions and move to the next problem. Feedback for correctness was not provided.

The procedure was the same for both conditions. Participants first completed a pretest and were introduced to the problem-solving interface. They then solved 12 problems, which were each paired with one example (experimental phase) and as the final step, completed a post-test.

4.2.2 Key Results from Jennings and Muldner (2020)

The aim of the original study was to investigate the effect of different assistance mechanisms (*fade in* vs. *fade out*) on learning. Learning was operationalized through gains from pre-test to post-test. The results, illustrated in Figure 4.3 b, showed that the *fade in* group learned more compared to the *fade out* group (p = 0.01).

Figure 4.2: *Algebra Tutorial Problem Solving Interface*

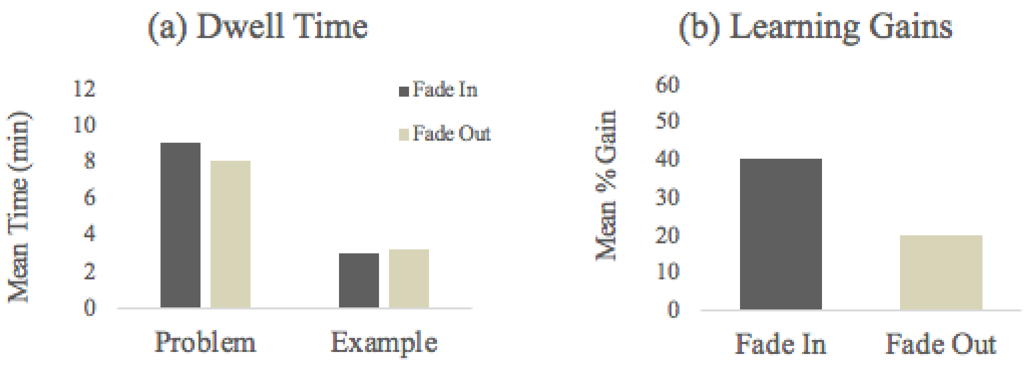


```
Example: Solve for a
b=((c(a+f))/d)-e
b+e=c(a+f)/d
d(b+e)=c(a+f)
(d(b+e))/c=a+f
((d(b+e))/c)-f=a
```

```
Problem: Solve for x
y=t((x-b)/z)+a
_
```

Note: Problem is on the right (area of interest, or AOI, shown in yellow) and the worked-out example is to the left (AOI shown in green). The AOIs were not visible to participants. A virtual keyboard, bottom, was used to enter the problem solution.

Figure 4.3: *Results from Jennings and Muldner (2020)*



Note: Graph (a) illustrates average total dwell time for each AOI and condition. Graph (b) illustrates average learning gains from pre to post test for each condition.

The original results also included basic analysis of eye tracking data, corresponding to dwell time in regions of interest. Two areas of interest (AOIs) were created within the problem-solving interface, namely the problem area and example area (see Figure 4.2). Dwell time in each AOI was calculated for every problem-example pair. Participants in the *fade in* condition spent significantly longer looking at the problem area compared to participants in the *fade out* condition (p = 0.049). There was no difference, however, in mean dwell time on the example AOI between the two conditions (see Figure 4.3 a).

In summary, participants in *fade in* condition had higher learning and devoted more dwell time to the problem area than the *fade out* condition. The original study also involved qualitative analysis of copying and correctness for a subset of the data by a human coder (without using eye tracking data). This analysis showed that both copying and correctness were affected by the type of assistance provided. While details are in the original paper, the *fade in* condition initially had lower copying and lower correctness than the *fade out* condition. This suggests that students in the *fade in* condition, which provided reduced assistance up front, struggled more at the beginning of the problem-solving session, possibly because they were trying to solve the problem without copying. Since copying reduces active processing, this may explain why this condition had higher learning. Importantly, the qualitative analysis in Jennings and Muldner (2020) suggested that students were using different strategies in the two conditions, copying vs. solving the problem on their own.

4.3 Present Scanpath Analysis

The goal of the current analysis was to further analyze the eye tracking data to examine how scanpaths were affected by the assistance mechanisms in the original study. As noted above, Jennings and Muldner (2020) used qualitative analysis on a subset of the data to provide initial evidence that different assistance mechanisms influenced students' strategies, namely whether they copied or solved the problem on their own. These strategies should affect the distribution of visual attention. If we can demonstrate scanpaths are different between the two conditions, this provides indirect support for difference in strategies as well. We had the following three research questions for analysis 2:

- Does problem-example similarity impact scanpath similarity?
- Does the type of assistance (*fade in* vs. *fade out*) impact scanpaths over time?
- Do learners settle into a strategy (copying vs. solving without copying) over time?

Before presenting the results related to these three research questions, we describe the data preparation and analysis framework.

4.3.1 Data Preparation and Analysis Framework

Data preparation. Recall that each participant solved 12 problems, with access to one example per problem, while an eye tracker logged their eye fixations. We extracted every participant's scanpath for each problem-example pair. Just as for the data preparation phase for analysis 1 (Chapter 3), each scanpath was stored in a csv file and included a series of fixations, namely the x and y coordinates for each fixation, as well as the duration of each fixation (in seconds). Only fixations that fell within bounds of the two target areas of interest (AOIs) were included in a scanpath (i.e., problem and example AOIs). Scanpath length was capped at 500 fixations to make the analysis feasible; however, the majority of scanpaths were shorter. Due to requirements of the MultiMatch tool, any scanpath shorter than three fixations was eliminated.

The MultiMatch tool includes the option to group fixations within the scanpath, which simplifies the analysis, something that is important to consider given the longer scanpaths in the present analysis. For this analysis, we followed the suggested thresholds to group successive saccades with amplitudes shorter than 40 pixels, group intermediate fixations shorter than 0.1 seconds in duration, and group angle change from one fixation to another less than 10% of the screen diagonal (for the present data, this correspond to 220 degrees) (Jarodzka et al., 2010; Holmqvist et al., 2011; Dewhurst et al., 2012). MultiMatch also requires the dimension of the computer screen used for the visual presentation to be input - this was set to 1920 x 1080.

Analysis framework. We followed the methodology from prior work (Dewhurst et al., 2018; Zhou et al., 2016) to use MultiMatch to compare all possible pairs of scanpaths within each condition and across the conditions (referred to as within-condition and across-condition comparison below). The within-condition comparison was the same as for analysis 1 (Chapter 3), except that scanpaths in analysis 2 came from different participants (due to between-subjects design of the study) and were only compared for a given problem-example pairing. To illustrate, for a given problem-example pair in a given condition, participant 1's scanpath was compared to participant 2's scanpath, participant 1's scanpath to participant 3's scanpath, participant 1's scanpath to participant 4's scanpath, and so on. The within-condition comparison was computed for each problem-example pair in the *fade in* and *fade out* conditions, each trial producing two collections of scanpath similarity scores, one per condition. The across-condition procedure was similar to that of Chapter 3 (i.e., for a given problem-example pair, every scanpath in one condition was compared against every scanpath in the other condition). Each comparison produced a similarity score for each of the five MultiMatch features (shape, direction, position, length and duration). We now describe how we applied this analysis framework to answer each of the three research questions.

(1) Does problem-example similarity impact scanpath similarity?

To address question 1 about the impact of problem-example similarity on scanpaths, we examined scanpaths corresponding to the first problem-example pair. Recall that for this problem, participants in the *fade in* condition were presented with an example aimed to discourage copying because its similarity to the corresponding problem was reduced, while participants in the *fade out* condition were given a highly-similar example facilitating copying. The within-condition comparison was computed twice, once for all scanpaths from the *fade in* condition, and once for all scanpaths from the *fade out* condition. Research question 1 also included the across-condition comparison. This produced three groups of MultiMatch similarity scores for each feature, which we refer to as $\text{Similarity}_{\text{fade in}}$, $\text{Similarity}_{\text{fade out}}$, and $\text{Similarity}_{\text{across}}$.

The similarity scores were analyzed using inferential statistics. As we stated for analysis 1, if the two within-condition similarity scores, $\text{Similarity}_{\text{fade in}}$, $\text{Similiarity}_{\text{fade out}}$, were higher than the $\text{Similarity}_{\text{across}}$ scores, this suggests that problem-example similarity differentially impacts cognitive processes in the two conditions. Moreover, if there was a difference between $\text{Similarity}_{\text{fade in}}$ and $\text{Similarityscores}_{\text{fade out}}$ scores, this would be evidence that problem-example similarity impacts the consistency of scanpaths.

(2) Does the type of assistance (*fade in* vs *fade out*) impact scanpaths over time?

Recall that in the original study, the *fade in* condition initially presented participants with reduced assistance, whereas the *fade out* condition started with high assistance; in both conditions, the type of assistance changed over time. For question 2, our goal was to analyze whether scanpaths later in the instructional session were affected by assistance provided earlier in the session, i.e. to test if scanpaths changed over time in each condition. By doing so, we aimed to indirectly assess the change in solving strategies between conditions.

To realize this goal, we needed to choose two points in the instructional sequence that would allow us to measure change in scanpath similarity over the instructional sequence. However, we also had to control for the effect of problem-example similarity, which differed between the two conditions for any given problem (see Figure 4.1). We selected two problem-example pairs that addressed these constraints: the first pair and the ninth pair in the 12-pair instructional sequence (see Figure 4.1). The two pairs had the same problem-example similarity within a given condition (high assistance in *fade out* condition and reduced assistance in *fade in* condition). Thus, the first problem-example pair served as a baseline allowing us to control for problem-example similarity and in doing so, isolate the impact of assistance over time. We already had data for the first pair from the analysis done for question 1. For the ninth problem-example pair, we followed the same procedure to compute the within-condition comparison. Using the results from question 1 related to pair 1, we then calculated scanpath similarity difference scores in each condition by matching comparison pairs between pair 1 and pair 9. A comparison pair consisted of two participants whose scanpaths were compared (e.g., participant 1 and participant 2). For each matched comparison pair, the scanpath similarity score at problem 1 was subtracted from the scanpath similarity score at pair 9. This produced two groups of similarity difference scores, one for the *fade in* condition and one for the *fade out* condition. The across-condition comparison was not included in the analysis for question 2 because the goal was to analyse whether there were differences in how much scanpaths changed over time rather than the way in which these changes occurred.

The final step for question 2 involved the analysis of the scanpath similarity difference scores using inferential statistics. Within a condition, if there was little difference in similarity scores between pair 1 and pair 9, this indicates that visual patterns did not change much over time. If the difference scores were positive, participants within a condition were converging on a similar viewing pattern by pair 9; if difference scores were negative, viewing patterns in a condition become more variable over time.

(3) Do learners settle into a strategy (copying vs. solving without copying) over time?

The third research question is a follow up to question 2, where we compare scanpaths at three points over the course of the learning session to analyze if there are changes in scanpath similarity over time via more complex trends over time than for question 2. We selected three problem-example pairs to represent visual patterns at the beginning, middle, and end points of the instructional sequence. As described above, in order to control for problem-example similarity, this similarity must remain the same for all three problems within a given condition. To account for this, problem-example pairs 1 and 9 were used as the start and end points for the analysis and pair 5 was chosen as the middle point. Pair 5 is the first problem participants encounter after a experiencing a shift in problem-example similarity at problem 4 (see Figure 4.1).

We already had data for the first and ninth pair from the analysis done for question 2. The procedure used to calculate similarity scores for pair 1 and 9 (within-condition comparison) was also used for problem-example pair 5. The corresponding scanpath similarity scores were analyzed using inferential statistics. As was the case for question 2, an across-condition score was not computed for this question as doing so would not help address the research question.

4.3.2 Results

We used MultiMatch to compute similarity scores for scanpaths for each of the three questions using the process described in the above. Each scanpath comparison resulted in five similarity scores: shape, direction, length, position and duration. Recall that a given similarity score ranges from 0 to 1, with higher scores representing higher scanpath similarity.

Table 4.1: Descriptives for Analysis 2 Similarity Scores

Descriptives (mean, SD for MultiMatch Features for problems 1, 5 and 9)

	Shape	Direction	Length	Position	Duration
Problem 1					
fade in	.9873 (.002)	.7506 (.066)	.9851 (.003)	.9163 (.039)	.6627 (.032)
fade out	.9847 (.003)	.7765 (.037)	.9843 (.003)	.9107 (.038)	.6676 (.032)
Problem 5					
fade in	.9875 (.001)	.7915 (.034)	.9864 (.002)	.9240 (.0237)	.6659 (.035)
fade out	.9841 (.003)	.7934 (.032)	.9828 (.004)	.9018 (.0387)	.6754 (.036)
Problem 9					
fade in	.9866 (.002)	.7374 (.126)	.9842 (.004)	.9181 (.035)	.6738 (.040)
fade out	.9840 (.003)	.7865 (.043)	.9818 (.005)	.8947 (.046)	.6721 (.037)

Question 1. We begin with the results for question 1 on the impact of problem-example similarity on scanpath similarity. As described above, question 1 included scanpaths from pair 1 only, where the *fade in* condition presented a reduced similarity problem-example pair and the *fade out* condition presented a high similarity problem-example pair. We obtained similarity scores that came from comparing scanpaths within the *fade in* condition ($\text{Similarity}_{\text{fade in}}$), within *fade out* condition ($\text{Similarity}_{\text{fade out}}$) and across the two conditions ($\text{Similarity}_{\text{across}}$). For each of the five MultiMatch features, we conducted a separate one-way ANOVA with comparison type as the 3 level between-subjects factor ($\text{Similarity}_{\text{fade in}}$,

Similarity$_{\text{fade out}}$ and Similarity$_{\text{across}}$) and similarity scores for the target feature as the dependent variable. Significant and marginally significant main effects were followed up with post-hoc Bonferroni comparisons.

The descriptives can be found in Table 4.1. The direction feature had a significant main effect, $F(2, 663) = 10.11, p < 0.01$. For the within-condition comparisons for this feature, Similarity$_{\text{fade out}}$ was significantly higher than Similarity$_{\text{fade in}}$ (p < 0.01) (see Figure 4.5). For the across-condition comparisons, Similarity$_{\text{fade out}}$ had significantly higher scanpath similarity than Similarity$_{\text{across}}$ ($p < 0.01$) but the difference between Similarity$_{\text{fade in}}$ and Similarity$_{\text{across}}$ was not significant.

The shape feature also had a significant main effect, $F(2, 663) = 50.72$, $p <$ 0.01, but in this case, Similarity$_{\text{fade in}}$ had significantly higher similarity scores than Similarity$_{\text{fade out}}$ ($p < 0.001$) (see Figure 4.4). For across-condition comparisons; Similarity$_{\text{fade in}}$ was significantly higher than Similarity$_{\text{across}}$ ($p < 0.01$) and Similarity$_{\text{fade out}}$ was significantly lower than Similarity$_{\text{across}}$ (p < 0.01). The other main effects were not significant (length, $F(2, 663) = 2.035, p = 0.131$; position, $F(2, 663) = 1.042, p = 0.353$; and duration, $F(2, 663) = 0.807, p = 0.446$.

Question 2. We now report on the results for question 2, which focused on the change in scanpath similarity from problem-example pair 1 to pair 9 in the instructional sequence. By the time problem 9 was encountered, participants had experienced the effect of type of assistance (*fade in* vs. *fade out*).

As noted above, because we used pair 1 as the baseline, we ran the analysis on the similarity difference scores (pair 9 – pair 1) for each scanpath comparison. We had two groups of scores for analysis 2 corresponding to scanpath difference scores within each of the two conditions: Similarity$_{\text{fade in}}$ and Similarity$_{\text{fade out}}$. Thus, for each of the five MultiMatch features, we conducted a separate one-way ANOVA with comparison type as the 2 level between subjects factor (Similarity$_{\text{fade in}}$ and Similarity$_{\text{fade out}}$) and difference scanpath similarity

Figure 4.4: *Mean Similarity Scores for Shape Feature (error bars represent 95% confidence intervals)*

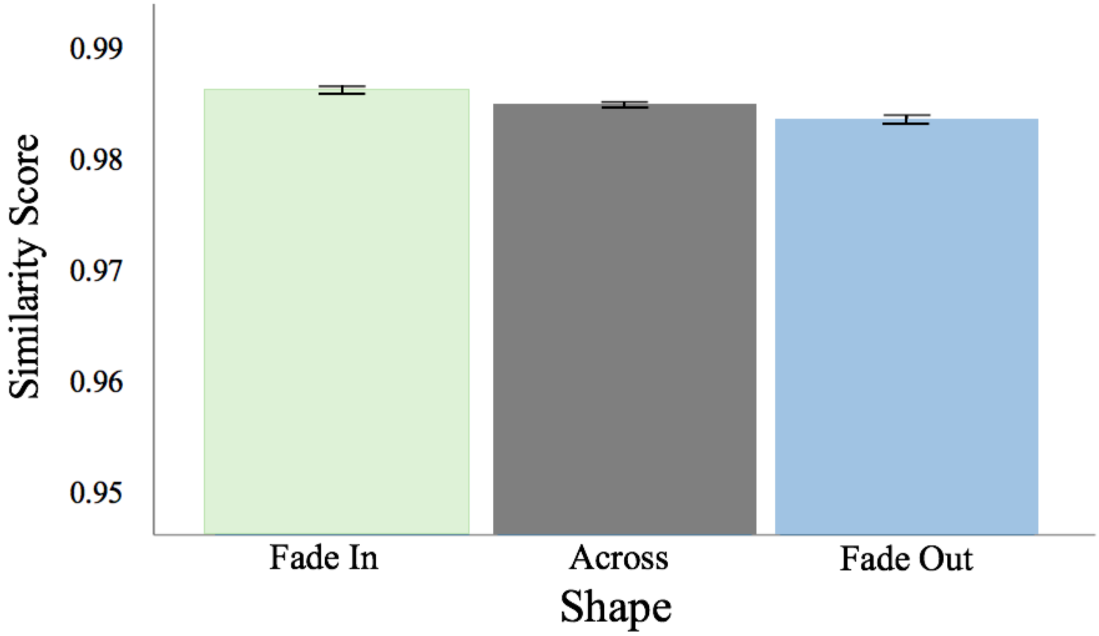


Figure 4.5: *Mean Similarity Scores for Direction Feature (error bars represent 95% confidence intervals)*

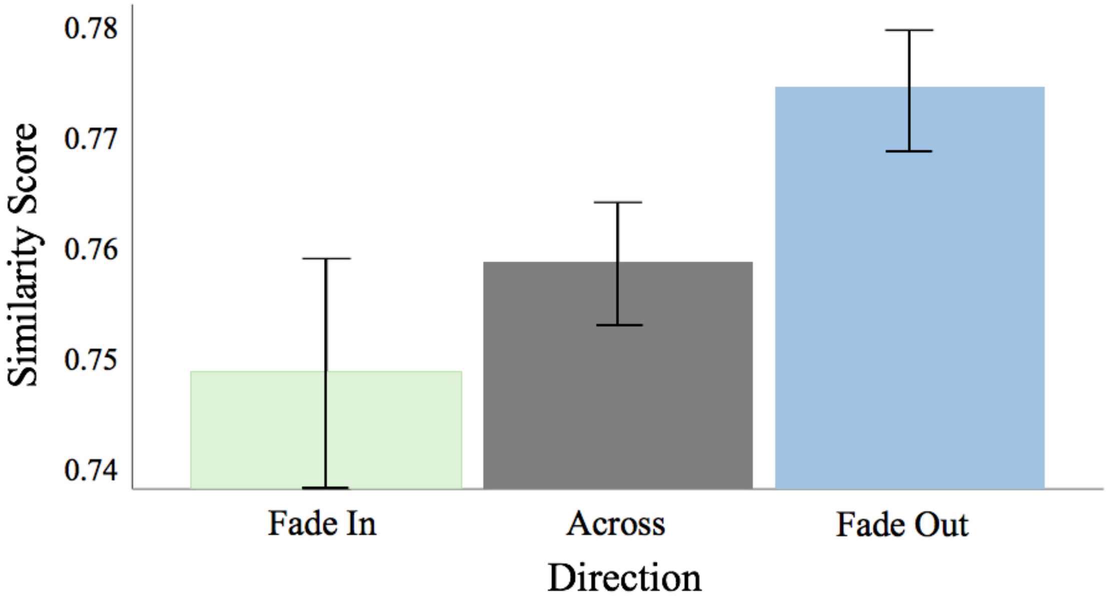


Figure 4.6: *Mean Similarity Scores for Length Feature (error bars represent 95% confidence intervals)*

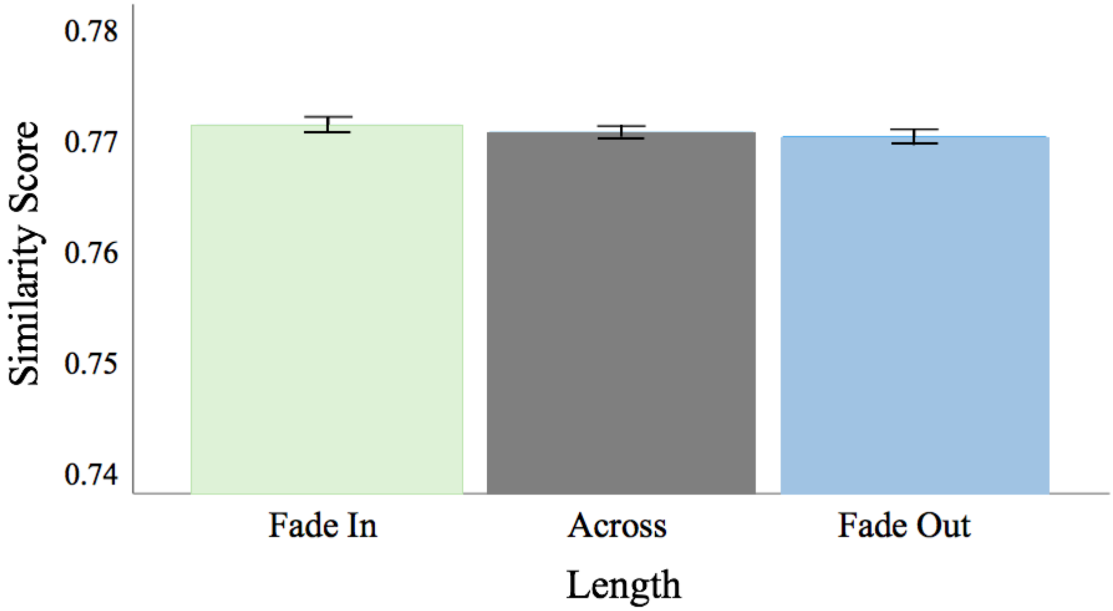
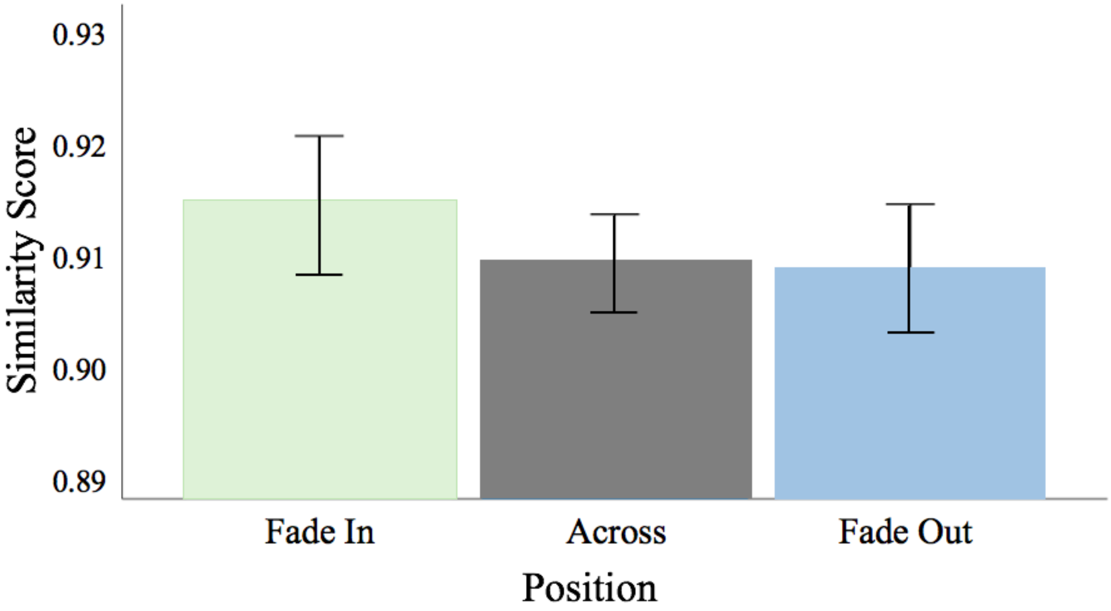


Figure 4.7: *Mean Similarity Scores for Position Feature (error bars represent 95% confidence intervals)*



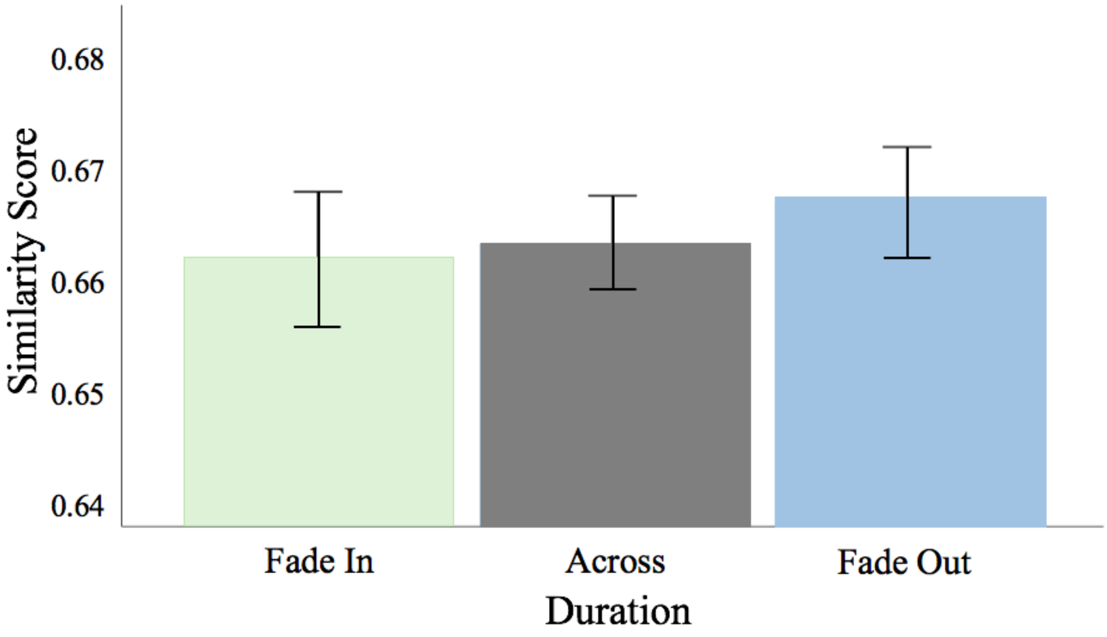
scores for the target feature as the dependent variable[2].

The descriptives for pair 1 and pair 9 can be found in Table 4.1. There was a significant main effect of comparison type on scanpath similarity for three MultiMatch features (length: $F(1, 304) = 7, p < 0.01$; position: $F(1, 304) = 18.2, p < 0.01$; and duration: $F(1, 304) = 8.1, p = 0.03$). For length and position (see Figure 4.9), $\text{Similarity}_{\text{fade in}}$ was significantly higher than $\text{Similarity}_{\text{fade out}}$, $p < 0.01$. In contrast, for duration (see Figure 4.9), $\text{Similarity}_{\text{fade out}}$ was significantly higher than $\text{Similarity}_{\text{fade in}}$, $p < 0.009$. The other main effects were not significant (direction: $F(1, 304) = 4.8, p = 0.029$; duration: $F\ (1, 304) = 8.1, p = 0.005$).

Question 3. We now report the results for question 3, which addressed the change in scanpath similarity over three points in the instructional sequence.

[2] Alternatively, this analysis can be run as a two-way ANOVA with the comparison group as between subject factor and time (problem-example pair 1, 5, or 9) as within subject factor, i.e., because the comparison pairs were matched between each time point.

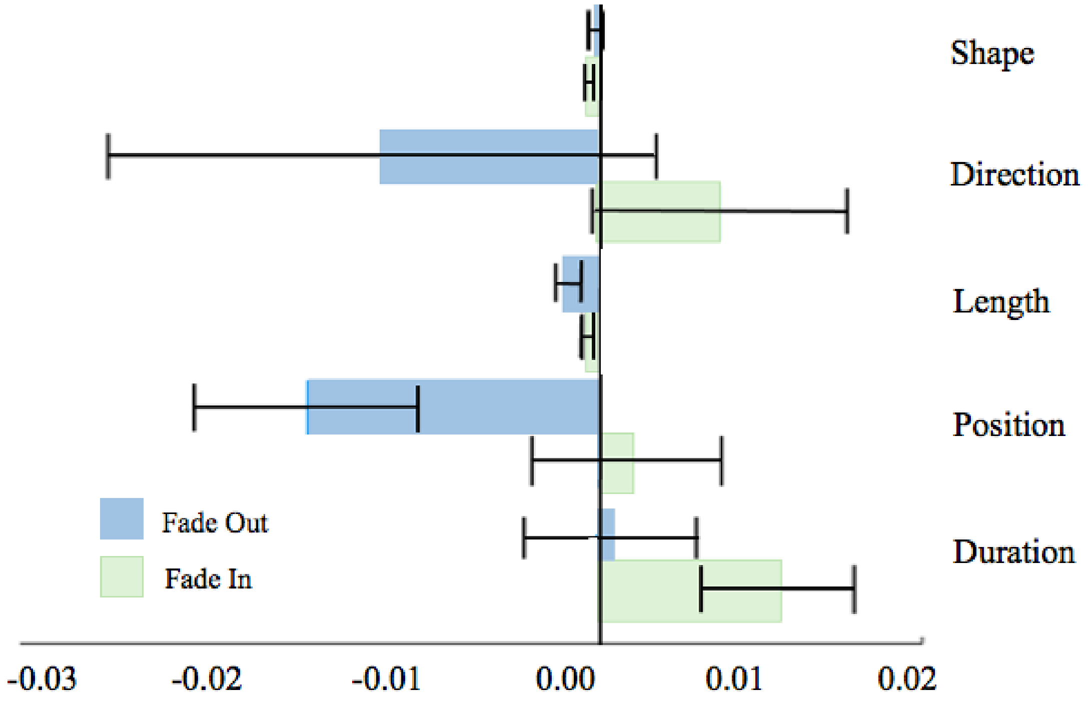
Figure 4.8: *Mean Similarity Scores for Duration Feature (error bars represent 95% confidence intervals)*



Here, we analyzed differences in scanpath similarity for problem-example pairs 1, pair 5 and pair 9, as that would let us investigate the change in consistency throughout the instructional sequence. In other words, three within condition groups were calculated, one at each problem-example pair (1, 5 and 9). For each of the five MultiMatch feature”s, we conducted a separate mixed two-way ANOVA with *problem number* as the 3 level within-subjects factor (problem 1, problem 5, problem 9) and *comparison group* as the 2-level between-subjects factor (Similarity$_{\text{fade in}}$ and Similarity$_{\text{fade out}}$).

The descriptives are in Table 4.1. There was a significant interaction between time and comparison group for all five features and since these are of primary interest, main effects will not be discussed (for the sake of completeness, they are in Appendix A). The results for each feature will be described individually

Figure 4.9: *Difference Similarity Scores Between Problem 1 and 9 (error bars represent 95% confidence intervals)*



below. [3]

For the shape feature, there was a significant interaction between time and comparison group, $F(2, 151) = 6.9$, $p < 0.01$. As shown in Figure 4.10, there is a relatively consistent level of scanpath similarity over time for both $\text{Similarity}_{\text{fade in}}$ and $\text{Similarity}_{\text{fade out}}$.

For the direction feature, there was a significant interaction between time and comparison group $F(2,\ 151) = 13.3$, $p < 0.01$. As shown in Figure 4.11, $\text{Similarity}_{\text{fade out}}$ scores remained relatively stable over time. In contrast, $\text{Similarity}_{\text{fade in}}$ for pair 1 and pair 9 was lower than $\text{Similarity}_{\text{fade out}}$, but at pair 5 similarity sharply increased to match similarity scores of $\text{Similarity}_{\text{fade out}}$.

For the length feature, there was a significant interaction between time and comparison group $F(2,\ 151) = 15.4$, $p < 0.01$. As shown in Figure 4.12, while $\text{Similarity}_{\text{fade in}}$ shows an increase in scanpath similarity at pair 5, $\text{Similarity}_{\text{fade out}}$ decreased.

For the position feature, there was a significant interaction between time and comparison group $F(2,\ 151) = 17.8$, $p < 0.01$. While $\text{Similarity}_{\text{fade out}}$ steadily decreased in scanpath similarity over time, $\text{Similarity}_{\text{fade in}}$ sharply increased at pair 5, compared to scores at pair 1 and pair 9 (as shown in Figure 4.13).

For the duration feature, there was a significant interaction between time and comparison group $F(2, 151) = 4.7$, $p = 0.01$. While $\text{Similarity}_{\text{fade in}}$ increased over time (see Figure 4.14), $\text{Similarity}_{\text{fade out}}$ increased only at pair 5, compared to the relatively similar scores at pair 1 and pair 9.

[3] The confidence intervals for the graphs in Figures 4.10, 4.11, 4.12, 4.13 and 4.14 were calculated using the method advocated in Field (2013) for repeated measure designs – the calculations were done within a given comparison group for each of the three problems.

Figure 4.10: *Within Condition Similarity Scores Over Time for Shape Feature (error bars represent 95% confidence intervals)*

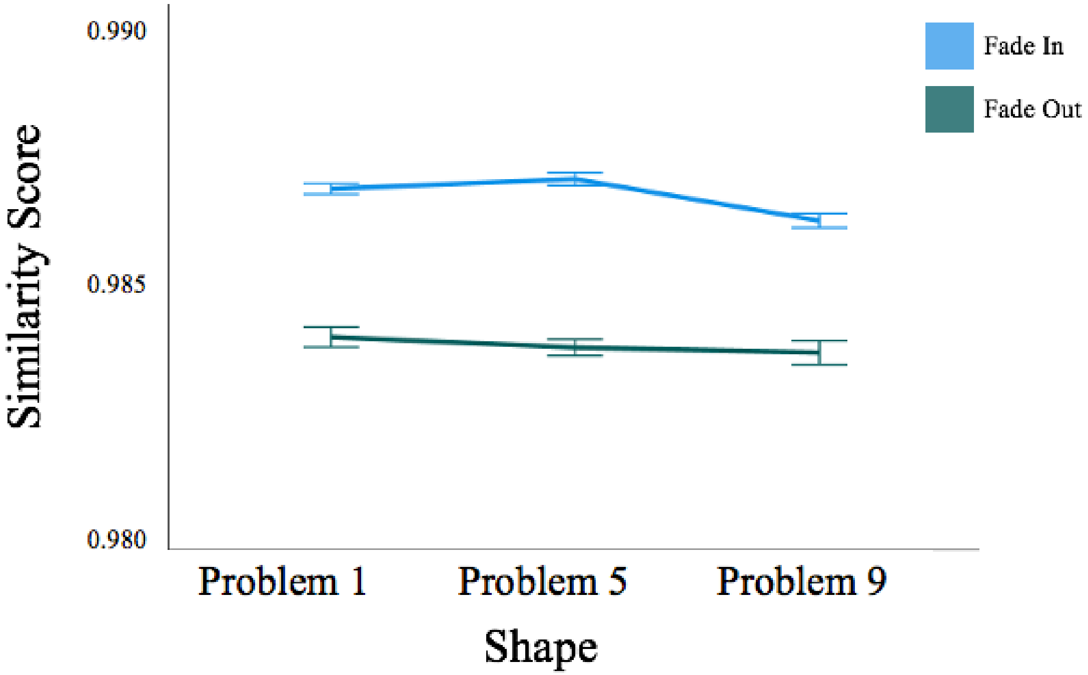


Figure 4.11: *Within Condition Similarity Scores Over Time for Direction Feature (error bars represent 95% confidence intervals)*

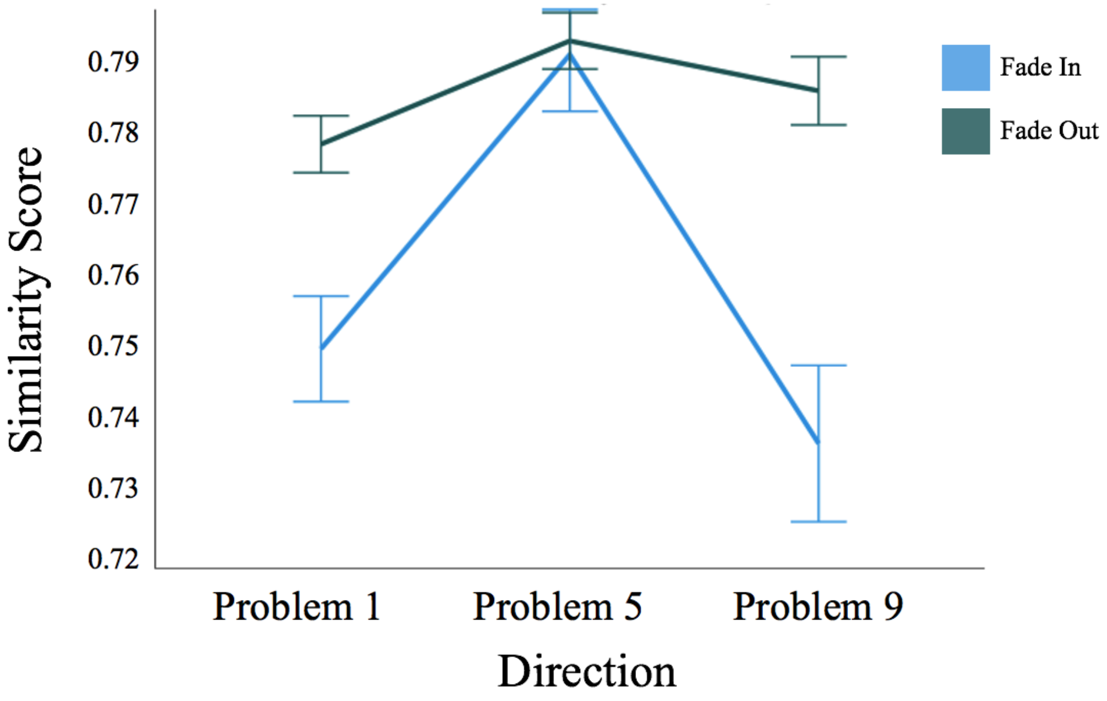


Figure 4.12: *Within Condition Similarity Scores Over Time for Length Feature (error bars represent 95% confidence intervals)*

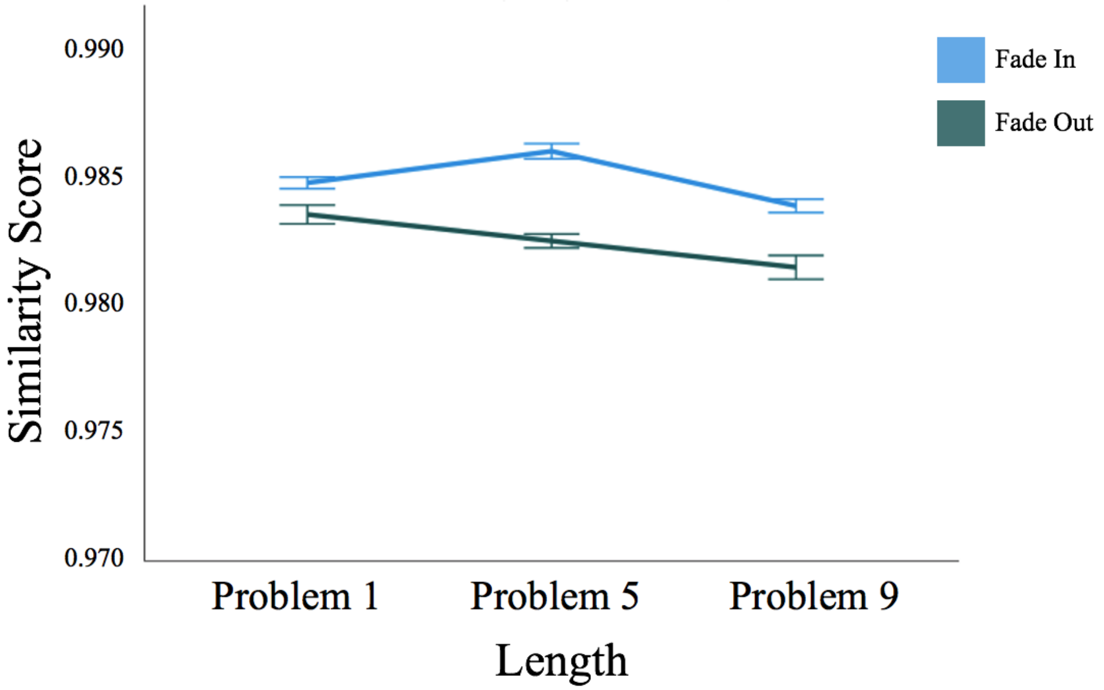


Figure 4.13: *Within Condition Similarity Scores Over Time for Position Feature (error bars represent 95% confidence intervals)*

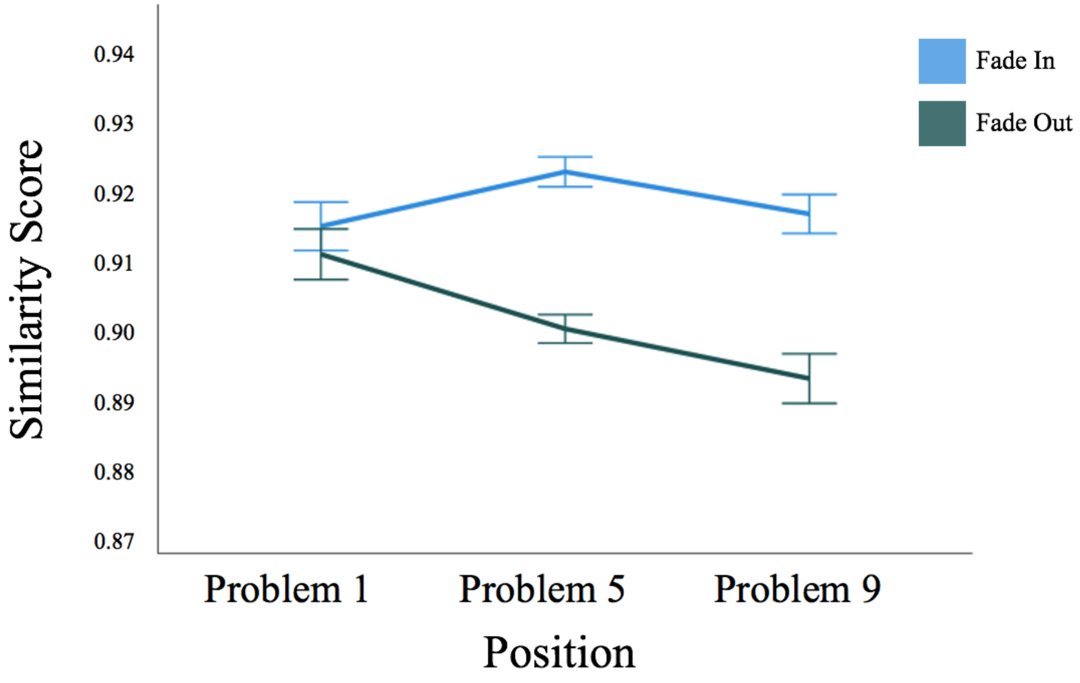
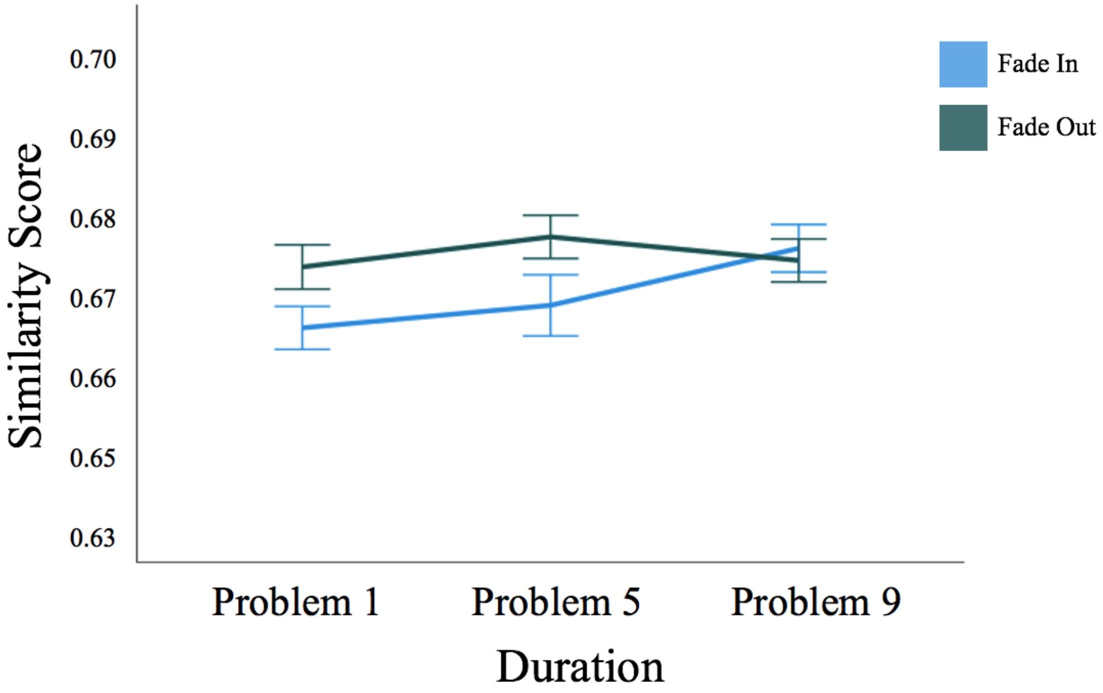


Figure 4.14: *Within Condition Similarity Scores Over Time for Duration Feature (error bars represent 95% confidence intervals)*



4.3.3 General Summary

In summary, difference in problem-example similarity impacted consistent use of scanpath strategy. However, for question 1, the across-condition similarity scores fell between the two within-condition scores, making interpretation difficult. We return to this issue in the final discussion.

5 Scanpath Simulation

5.1 Introduction

For both analyses in Chapters 3 and 4, we reported relationships between the within and across collections of similarity scores. This is the standard way of reporting results from scanapth analyses (Dewhurst et al., 2018; Zhou et al., 2016) but existing sources provide little guidance beyond stating that if conditions elicit different cognitive processes, then the across-condition similarity score should be lower than within-condition similarity scores (Zhou et al., 2016; Mathôt et al., 2012). However, interpretations for other relative orderings of the similarity scores are not included (e.g., what is the implication of results when the across-condition mean similarity score falls in between the two within-condition scores, as was the case for some of our results?).

To address this gap, we implemented a simulation of scanpath comparisons, which mimics at a high level an experimental study by including two 'study conditions' and compares 'scanpaths' within and across the 'conditions'. We put study conditions and scanpaths in quotes because these are abstractions of the experimental constructs, as we describe shortly below. The goal of the simulation is to provide guidelines for interpretating results from a standard scanpath analysis framework, i.e., the within- and across-group similarity scores. The relative orderings of the similarity scores are impacted by the proportion of viewing patterns in each condition but to date how that occurs is not clear. What do we mean by viewing pattern? A viewing pattern produces a scanpath – this pattern is influenced by the strategies participants use when interacting with instructional materials, for instance, to solve problems. Thus, strategies produce scanpaths, and for the purposes of the simulation, these terms are synonymous.

To analyse the impact that relative proportion of strategies in a given condition has on scanpath similarity scores, the simulation makes the simplifying assumption that there are only two 'strategies' in the simulated experiments (i.e., types of scanpaths), referred to as strategy1 and strategy2 (e.g., participants either solve a basic problem through retrieval or through recasting to a different format first). In reality, it is possible that participants use more than two strategies in a given experiment. Here, however, we made the simplifying assumption related to the number of strategies to make the simulation feasible but also because the experiments analyzed in this thesis are each hypothesized to have two solving strategies (mediated retrieval vs. direct retrieval; copying vs. solving without copying).

We used the simulation to investigate several scenarios, in which the proportion of each strategy was varied to observe the change in similarity scores for within and across group comparisons. Below, we provide further details on the simulation set up.

To implement the simulation, we created a Python program that uses two lists, List1 and List2. These two lists represent two hypothetical study conditions. In Python, a list is a sequence of elements, which here we call 'items'. In our simulation, each item represents a single scanpath. However, in the simulation we do not need to encode actual scanpaths because we can abstract that information by assigning a simple label to each list item to indicate whether it represents strategy1 or strategy2 (recall that the simulation assumes the presence of two strategies). We can vary the proportion of labels in a given list (e.g., populate List1 with 80% strategy1 and 20% strategy2 items). Note that this represents the proportion of a strategy in a given condition (e.g., 80% scanpaths came from people who copied and 20% from people who solved without copying). Alternatively, we could set 100% of the labels in a list to the same label, which would represent the scenario in which all participants used the same strategy.

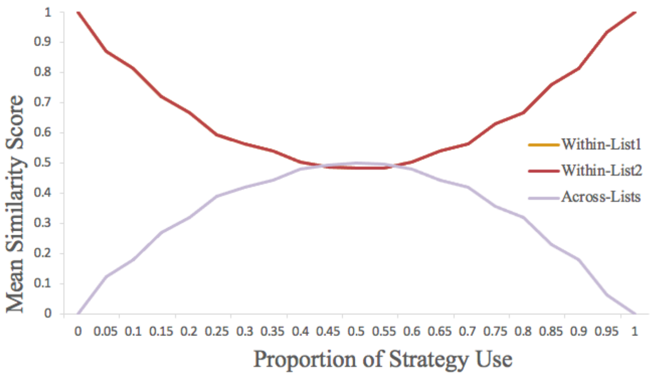
In the simulation, each list consists of 30 items, as this was approximately the same number of scanpaths per participant (in Tan, Muldner, and LeFevre (2016) study) or per trial (in Jennings and Muldner (2020) study). To summarize, the simulation consists of two lists (List1 and List2, representing two conditions), each containing 30 items (scanpaths). The simulation code then computes three mean similarity scores, using the method used in our actual scanpath analysis, summarized below:

- Within-List1: each item in List1 is compared to each other item in List1 in a pairwise fashion. Every comparison results in a similarity score. For our initial simulations, we set t result of this comparison to be 1 if both items are labelled as the same strategy (strategy1 and strategy2 or strategy2 and strategy2) and 0 otherwise. We also subsequently investigated the effect of changing the similarity score. The similarity scores are tallied and the mean within-List1 score is computed.

- Within-List2: the same process as for List1 except that List2 items are involved.

- Across-List: each item from List1 is compared to each item from List2 in a pairwise fashion, similarity scoring for matches is strategy type (strategy1 or strategy2) is the same as in within-list comparisons. Similarity scores are tallied and the mean across-List score is computed.

Figure 5.1: *Changes in Similarity Scores as Both Lists Shift from Use of One Strategy to Another Strategy*



Note: The X-axis shows the proportion of one strategy to the other; for List1, this indicates the proportion of strategy1, while for List2 this indicates proportion of strategy2. (e.g., when x = .2, List1 has 20% strategy1 and List2 has 20% strategy2).

The program writes the three mean similarity scores (within-List1, within-List2 and across-Lists) to an excel csv file. The simulation iteratively manipulates the proportion of items (i.e.. strategies) in each list to determine the effect on the three comparison scores – thus, each simulation involved a loop, where one iteration corresponded to a particular proportion setting. We then generated graphs from the csv files. In order to investigate the relative similarity of the three comparison groups and how they change given the full scope of different strategy proportions we ran four different case studies, presented below.

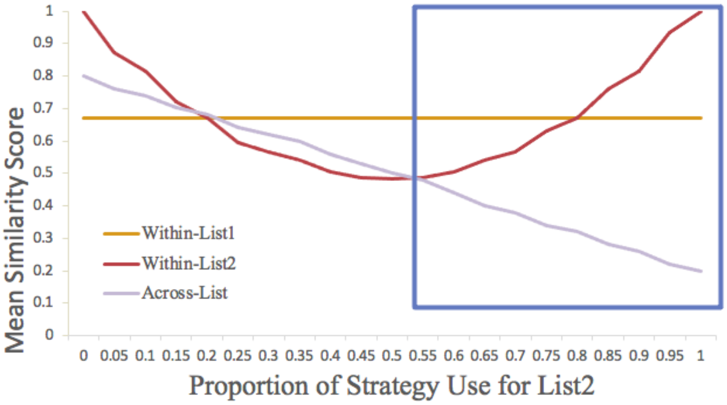
5.2 Case Study 1 – When two conditions have same proportion of viewing strategies

In case study 1, the proportion of strategies is iteratively varied in each of the two lists that represent the two study conditions. This is done to illustrate the impact of consistency of a given strategy within each list on the across-list similarity scores. Initially List1 is comprised only of strategy2 items and List2 only of strategy1 items. In the second iteration of the simulation, the percentage of the other strategy, i.e., one that is not already present in the list, is increased by 5% (i.e., in the second iteration, List1 is composed of 5% strategy1 and 95% strategy2 items); In the third iteration, List1 is composed of 10% strategy2 items and 90% strategy1 items, and so on. This continues until the final iteration of the simulation, where each list is now composed entirely of the strategy items that were not initially present in the list. At each iteration of the simulation, the program calculates three mean similarity scores: (1) the within-list similarity for List1, (2) the within-list similarity for List2, and (3) the across-list similarity for List1 and 2. The resulting similarity values are shown in Figure 5.1. The X-axis shows the proportion of the one strategy in the list vs. the other strategy , while the Y-axis shows the mean similarity score for each of the three comparison groups. There appears to be only one line in Figure 5.1 for the within-list scores because the same within-list similarity scores at each point in the simulation are the same. This may seem counterintuitive as the two lists differ in which strategy is most prevalent throughout the simulation but the two lists have same strategy proportions throughout the simulation, thus producing the same results in similarity score values.

As shown in Figure 5.1, when one strategy is much more common than the other strategy within a list, the within-list comparisons have high similarity scores. For instance, when List1 represented by the red line in Figure 5.1 has 95% of items labelled as strategy1, then the within-list similarity is 0.87. As presence of the two strategies in a given list evens out (becomes a 50:50 ratio), then the mean within-list similarity lowers to about 0.5 for both lists.

The across-list similarity scores are affected by the within-group scores. When the two lists are mainly composed of different strategies (e.g., List1 mostly has strategy1 and List2 mostly has strategy2), the across-list comparison has low similarity (e.g., see Figure 5.1, x = .95, here the across list similarity is very low, at 0.05). When the two lists have an equal proportion of the two strategies (50:50 ratio), the across-list similarity score increases, enough to overlap with the within-list similarity scores and even surpass them slightly. This result is due to the characteristics of the within versus across-group comparison set up. For within comparisons, all items are compared against the other items within that list; self-comparisons are not performed (i.e., item1 is never compared to itself). For the across-group comparisons all items in one list are compared to all items in the other list. Because the number of comparisons for each type of group (within vs. across) are different, this impacts final similarity score but this effect lessens as the total number of items in each list increases. The take-away is that across-condition comparisons will produce a low similarity score if different conditions elicit different viewing patterns (i.e., scanpaths capturing strategies), and the pattern within each condition is consistent (i.e., most of the scanpaths follow this pattern). On the other hand, across-condition scores will be similar to within-condition scores when the two conditions have similar strategy composition.

Figure 5.2: *Changes in Similarity Scores when List1 is Composed of 80% strategy1*



Note: The X-axis shows the proportion of strategy2 for List2, e.g., when x = 0, List2 has 100% strategy1, when x = 0.0.5, List2 has 95% strategy1, and so on until x = 1 and List2 has 0% strategy1 and 100% strategy2.

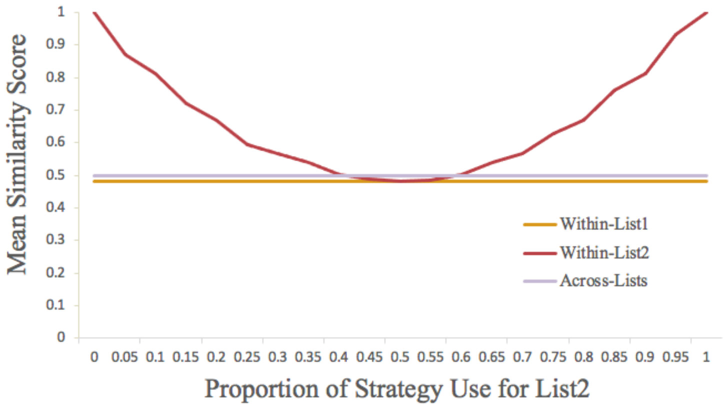
5.3 Case Study 2 – When two conditions have different proportions of viewing strategies

In case study 2, we investigate the effects on similarity scores when one list is made up of mostly one strategy while varying the proportion of the strategies in the other list. Specifically, we keep the proportion of the two strategies in List1 the same throughout the simulation – here these were set to 80% for strategy1 and 20% for strategy2. As illustrated by Figure 5.2, the same proportion of strategies within-List1 results in a straight line, indicating that as expected the mean within-list score does not change throughout the simulation (see orange line in Figure 5.2). List2, on the other hand, starts off as entirely composed of strategy 1 (see x = 0 in Figure 5.2), gradually shifting into being entirely composed of strategy2 (see x = 1 in Figure 5.2). This creates the same U-shaped curve as observed in case study 1, representing the similarity scores at each iteration of the simulation, with the high within-list similarity scores produced by iterations when one strategy is much more frequent and lower scores produced iterations where there is a more proportionate presence of the two strategies.

When List2 is composed 100% of strategy1 items and List1 is composed of 80% strategy1 items, the across-list similarity scores are high (see Figure 5.2 at x = 0, where across-list similarity score is 0.8). As the proportion of strategy1 items decreases in List2, the across-list similarity score also decreases (e.g., in Figure 5.2, at x = 0.05, see across-list similarity value compared to across-list similarity value at x = 0.3) until List2 is composed entirely of strategy2. Since List1 is made up of only 20% strategy2 items, this produces a low across-list similarity score of 0.2. Thus, the across-list similarity score is an important indicator of the proportion of strategies in each list. Below is a summary of the relationships between across- and within-list similarity scores and what they represent in regards to strategy prevalence in each list:

- Across-list similarity scores are only higher than within-list scores when at least one list has exactly equal presence of all viewing strategies.
- Across-list similarity scores overlap, or are the same as, within-list scores when both lists have same ratio of strategies (see x = 0.2 in Figure 5.2).
- The across-list similarity score falls in between the two within-list scores when both lists are mainly composed of the same strategy, but one has a higher proportion of that strategy.
- The across-list similarity score falls below both within-list similarity scores when the two lists are not mainly composed of the same strategy (see the blue square box in Figure 5.2). When this occurs, it indicates that the two

Figure 5.3: *Changes in Similarity Score when List1 is Composed of Equal Proportion of Both Strategies*



Note: The X-axis shows the proportion of strategy2 for List2, e.g., when x = 0, List2 has 100% strategy1, when x = 0.0.5, List2 has 95% strategy1, and so on until x = 1 and List2 has 0% strategy1 and 100% strategy2.

lists are composed of different strategies, even if both within-list scores are high (e.g., x = 0.8 in Figure 5.2). In experimental settings, this indicates there is an effect of condition that influences the viewing strategy.

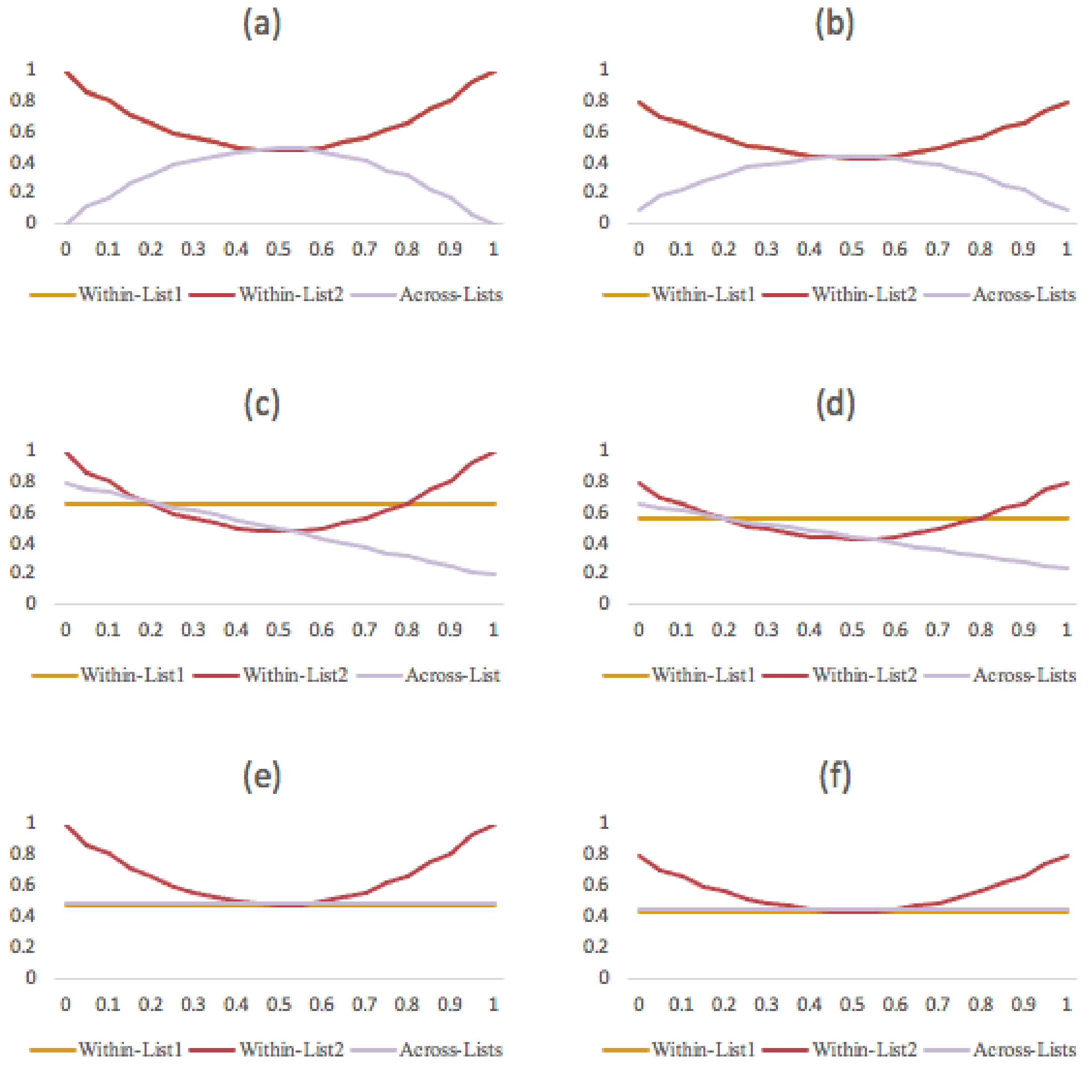
5.4 Case Study 3 – When no prevailing strategy exists in one condition

Case study 2 investigated what happens to the across-list similarity when one list is made up mostly of a single strategy while varying the proportion of strategies in the second list. Case study 3 has a similar set up to case study 2, except here we examine what happens when List1 has an equal proportion of the two strategies (i.e., 50% strategy1 and 50% strategy2, held constant across the simulation). As was the case study 2, the proportion of strategy1 vs. strategy 2 in List2 changes over the course of the simulation, shifting from being composed entirely of strategy1 to being composed entirely of strategy2 (see U-shaped curve for List2 in Figure 5.3).

The simulation results are in Figure 5.3. The shape of the across-list similarity graph (see gray line, Figure 5.3) is different from the one in case study 3. Instead of the mean across-list similarity score descending in value over time as the presence of strategy1 in List2 becomes less common than strategy2 (case study 2), the across-list similarity scores remain constant, with a mean similarity score of 0.5. Why is the across-list score not impacted by the change in proportion of the two strategies in List2? In case study 2, List1 had an 80:20 ratio for strategy1 to strategy2. This meant that when comparing items across lists, there were more matches when the dominant, more common strategy in List1 (strategy1) was also the dominant strategy in List2. In contrast, for case study 3, list 1 has no dominant strategy as both strategies are equally common 50:50). Thus, when comparing across lists, any given item from List2 will be compared against all items in List1 and whether the item is labelled strategy1 or 2, it will result in 50% matches (similarity score = 1) and 50% misses (similarity score = 0). This illustrates that across-condition scores are not a measure of overall consistency of both lists grouped together, but rather a measure of how alike one group is to the other.

The important take away from this scenario is that if one condition is composed of an equal proportion of the strategies, no matter the strategy proportions in the other condition, the across-condition similarity score will be 0.5.

Figure 5.4: *Change in Mean Similarity Score for Case Studies 1, 2 and 3 with Binary Scoring and Fuzzy Scoring*



Note: For graphs (a) and (b), the X-axis shows the proportion of strategy1 for List1 and strategy2 for List2 (e.g., when x = 0, List1 has 100% strategy2 and List2 has 100% strategy1). For graphs (c) - (f), the X-axis shows the proportion of strategy2 for List2.

5.5 Case Study 4 – Effect of 'Fuzzy' Scoring

Thus far in the simulations, similarity was measured in binary terms: a score of 1 was assigned to comparisons that had an exact match (strategy1 vs. strategy 1 or strategy2 vs. strategy2) 0 otherwise. In case study 4, we examine the effect of 'fuzzy' matches, to better capture the reality of scanpaths similarity measurements (i.e., in experimental results, scores were never 0 or 1, instead falling into a range between these extremes). To test the effect of fuzzy scoring, in the simulation code we changed the similarity scoring system so that when two item labels were compared against each other and matched, they scored 0.8 instead of 1; when two item labels were compared against each other and did not match, scores were set to 0.1 instead of 0. We then re-ran simulations for case studies 1, 2 and 3. The results are in Figure 5.4.

For each case study, as expected, the similarity scores have slightly lower values and a smaller magnitude of change over the course of the simulation. For instance, the across list range in case 2 for binary similarity scores ranges from 0.8 to 0.2 (graph c in Figure 5.4), but only from 0.6 to 0.25 in the fuzzy scoring simulation (graph d in Figure 5.4). Importantly, however, the interactions between within List1, within List2 and across-list similarity scores were not altered given this fuzzy scoring system and thus the conclusions drawn in case studies 1 to 3 hold. For instance, the finding from case study 1 holds in the fuzzy scoring simulation: when the two lists have the same proportions of strategies, the within-list scores present an overlapping U-shape and the across-list reflects that U shape (see graph (b) in Figure 5.4). When case study 2 was re-run with the fuzzy scoring system, the across-list scores still descended as List2 increased in strategy2 prevalence (see graph (d) in Figure 5.4). When List1 was composed of equal amounts strategy1 and strategy2, even with fuzzy scoring to mark similarity for the two lists, within-List1 and across-list still resulted in overlapping similarity scores (albeit these were slightly lower, 0.42 for fuzzy similarity see graph e, and 0.5 for binary similarity, see graph (f) in Figure 5.4).

The takeaway from case study 4 is that although the scoring system from simulations is simplified, the conclusions regarding prevalence of strategies within and across strategies should be generalizable to scanpath analyses that rely on fuzzy similarity scores.

6 Discussion

The goal of this thesis was to contribute to the field of scanpath investigation by examining whether scanpaths can distinguish between different instructional material formats. Here, we begin by discussing the results from 3 and 4 in light of the simulation findings. We also discuss further implications, limitations, and future work.

6.1 Analysis 1: Tan, Muldner, and LeFevre (2016) Data Set

6.1.1 Overview

The data set for analysis 1 came from a within-subjects study with two conditions in which participants were tasked with solving division problems that were presented in either a *traditional* or *recast* format. The *traditional* division format was hypothesized to elicit a mediated retrieval process (first transform the problems into multiplication format, then solve the recast problem). In contrast, the *recast* format was hypothesized to bypass the transformation step as the problem was already in multiplication format, leading the student to directly retrieve the solution from memory. The original analysis provided evidence for this hypothesis, because problem format affected visual attention and solving time (faster in the *recast* condition). Our analyses on similarity of scanpaths showed there were differences in visual patterns for three of the five MultiMatch features between the two problem formats.

6.1.2 Relationships between the Three Comparison Groups

The relationship between the comparison group similarity scores for the length and position MultiMatch features was the same from a descriptive standpoint (although not all pairwise comparisons were significant): $\text{Similarity}_{\text{Traditional}} > \text{Similarity}_{\text{Recast}} > \text{Similarity}_{\text{Across}}$ (see Figures 3.6 and 3.7). This relationship, where descriptively the across-condition score is lower than both the within-condition scores (though not all pairwise comparisons were significant), indicates that the two conditions elicit different viewing strategies (Mathôt et al., 2012). The interpretation is supported by our simulation results (see Figure 5.2, blue box area highlighted), where the mean across-List similarity score was only lower than both the within-list similarity scores when the two lists mainly contained two different strategies (i.e., dominant strategy use was different between the two lists).

Unlike the length and position features, for the shape feature the ordering of the three comparisons scores was different, because the across-condition score fell between the two within-condition scores ($\text{Similarity}_{\text{Traditional}} > \text{Similarity}_{\text{Across}} > \text{Similarity}_{\text{Recast}}$ (see Figure 3.4). Our simulation results suggest this relationship means that there is a high prevalence of the same viewing strategy in both conditions (see Figure 5.2).

The shape feature finding appears contradictory to the length and position feature findings. One potential explanation is that the measurement of similarity via vector shape may not be as sensitive in terms of capturing the different viewing patterns elicited by problem format.

6.1.3 Consistency of Viewing Patterns

For all three features (shape, position and length), the *traditional* format had higher mean similarity than the *recast* format. Higher similarity implies that people are consistent in how they view the instructional material (i.e. they tend to be using a particular viewing pattern). The higher viewing consistency for *traditional* presentation format was somewhat surprising. According to Tan, Muldner, and LeFevre (2016), the *recast* condition should elicit a direct recall of the answer, which we anticipated would result in more consistent viewing patterns (direct retrieval takes less time and requires less visual information to be stored in working memory (Huebner and LeFevre, 2018)). However, since division problems are usually solved through remediated retrieval for equations in the *traditional* format, then there might be the added element of familiarity. In contrast, a recast multiplication presentation may be a less familiar format, which would elicit varied viewing patterns (i.e., scanpaths).

6.2 Analysis 2: Jennings and Muldner (2020) Data Set

6.2.1 Overview

The data set for analysis 2 came from a between-subject study. Analysis 2 included two conditions from the original data set, the *fade out* condition (the similarity between problems and corresponding examples decreased over time) and the *fade in* condition (the similarity between problems and corresponding examples increased over time). The *fade out* condition was hypothesized to initially encourage copying and thus hinder learning. In contrast, the reduced similarity between problem and corresponding examples at the start of the *fade in* condition were hypothesized to challenge participants to produce the answer without copying and so encourage learning. We had three research questions for this data set, each discussed below.

6.2.2 Question 1 : Does problem-example similarity impact viewing patterns aka scanpath similarity?

There were significant differences between the within- condition and across-condition comparisons for two MultiMatch features. For the shape feature, $\text{Similarity}_{\text{Fade In}} > \text{Similarity}_{\text{Across}} > \text{Similarity}_{\text{Fade Out}}$ (see Figure 4.4). Our simulation results imply that a relatively high proportion of the same viewing strategy is being employed in both conditions, but that the *fade in* condition elicits more consistent use of a given viewing strategy compared to the *fade out* condition. For the direction feature, $\text{Similarity}_{\text{Fade Out}} > \text{Similarity}_{\text{Across}} > \text{Similarity}_{\text{Fade In}}$ (see Figure 4.5). This result aligns partly with the results for the shape feature, as it implies that the two conditions elicit a high proportion of the same strategy. However, here, it is the *fade out* condition that elicits more consistent use of the viewing strategy as supposed to the *fade in* condition.

Thus, the results for both features indicated that, for the first problem, both presentation formats elicited a higher proportion of a given viewing strategy. However, there were also conflicting results as to which condition elicited more consistent viewing patterns (i.e., which condition had higher mean similarity score).

The remaining two research questions did not involve the across-condition comparison, and so we will not be employing interpretations from simulation chapter in their discussion.

6.2.3 Question 2 : Does type of assistance (*Fade in* or *Fade out*) effect viewing patterns?

For question 2, we analysed difference scores, where similarity at problem 1 was subtracted from similarity at problem 9. Thus, a negative difference score indicates scanpath similarity decreased over time and positive difference scores indicates scanpath similarity increased over time. Two MultiMatch features showed considerable change between these two problems, namely position and duration. The duration feature showed a relatively large increase in similarity for the *fade in* condition over time, compared to a smaller increase in similarity for the *fade out* condition. The rise in scanpath similarity over time for the *fade in* condition may indicate that, after initial struggle to solve problems without example assistance, participants eventually may have learned to focus on relevant areas in the problem work space. This would increase the consistency of viewing patterns, as participants gravitated towards the same problem areas.

The position feature showed a substantial decrease in scanpath similarity for the *fade out* condition over time whereas the *fade in* condition showed a slight increase in similarity. When students were presented with highly similar example pairings at the beginning of problem solving in the *fade out* condition, this facilitated copying. When the problem-example similarity decreased as assistance was faded out over time, participants may have been less able to identify relevant areas of viewing to solve the problem, which may have resulted in more varied viewing patterns (i.e., scanpaths).

6.2.4 Question 3 : Do learners settle into a strategy (copying vs. solving without copying) over time?

For this question, we computed similarity within both conditions at three points in time: problem 1, 5 and 9 to observe shifts in consistency of viewing patterns over time. Three features, namely shape, position and length, showed similar patterns in change over time (see Figures 4.10, 4.13 and 4.12). Scanpath consistency for these three features decreased over time for the *fade out* condition (see section 6.3.2 for an explanation). In contrast, scanpaths in the *fade in* condition exhibited a peak in similarity at the middle of the instructional session, where problem-example pairs gradually become more similar. Perhaps this peak corresponded to an 'a-hah' moment of learning that directed participants' visual attention. These speculations require further investigation, such as qualitative data collection of participants' problem-solving processes.

The two remaining features, direction and duration, had different patterns in change of scanpath consistency over time. Noticeably, both features showed a peak in similarity for the *fade out* condition at the middle point of instructional sequence. We will discuss these results further in section 6.4.1.

6.3 Implications, Limitations and Future Work

6.3.1 Implications

As discussed in the above sections, our results show that to some extent scanpath analyses do distinguish different instructional material formats. Since these formats are hypothesized to elicit specific cognitive processes (described in their respective analysis Chapters 3 and 4), our findings provide indirect evidence that scanpaths distinguish cognitive processes.

Within and across condition scores demonstrated that for both data sets, presentation format affected viewing patterns captured by scanpaths, although scanpath similarity was impacted to differing degrees. The scanpaths from Tan, Muldner, and LeFevre (2016) data set elicited the expected lower across-condition scores compared to higher within-condition scores (as per Zhou et al. (2016) and Dewhurst et al. (2018)). These differences in comparison groups implies that scanpath similarity analysis can differentiate between conditions designed to elicit different cognitive processes. Prior results using alternative outcome measures also implied problem format impacted outcomes. Specifically, Tan, Muldner, and LeFevre (2016) found *recast* format showed increased fixation on middle operand and faster solving time compared to *traditional* format.

In contrast, the results from the Jennings and Muldner (2020) data set were less straightforward. For this data set, the across-condition mean scores fell between the two within-condition mean scores, complicating the interpretation of the results. Specifically, the interpretation from the scanpath simulations implies that while both conditions may elicit similar viewing patterns, the instructional material did impact the consistency of those patterns. Prior results also showed instructional material impacted alternative outcome measures. Jennings and Muldner (2020) found the *fade in* condition had higher performance scores than the *fade out* condition. Additionally, participants in *fade in* condition spent more time looking at the problem area compared to fixation results from the *fade out* condition.

These more complex outcome from scanpath analysis 2 is likely at least in part a function of task complexity. Dewhurst et al. (2018) found that increasing task complexity inherently reduces scanpath consistency. The algebraic problems in Jennings and Muldner (2020) study were more complex than the more basic division problems in Tan, Muldner, and LeFevre (2016) producing longer scanpaths and complicating the similarity analysis. Thus, MultiMatch may have been less able to identify viewing patterns. Moreover, the scanpath analysis for the Jennings and Muldner (2020) data set involved comparing scanpaths between participants, rather than within participants as was done for Tan, Muldner, and LeFevre (2016). Both approaches have been used in past work (Zhou et al., 2016; Dewhurst et al., 2018), and thus there is precedent for them. However, this factor may have also influenced scanpaths similarity.

These speculations, however, require further research. The present thesis is only a first step towards exploring scanpath methods in problem-solving tasks and so it has limitations and thus opportunities for future work.

6.3.2 MultiMatch Feature Interpretation

We chose the MultiMatch similarity measurement tool for our analysis in part due to its ability to provide similarity rating for five different scanpath features. Although this provides richer data on scanpath similarity, there is currently a lack of guidelines on how to interpret a given MultiMatch feature. This is particularly challenging when there are conflicting results for different features. For instance, direction and duration features demonstrated alternative scanpath similarity patterns over time (see question 3 in 4). It is worthwhile to note that the three features that produced similar results, namely shape, position and length, are all coordinate-oriented features (x, y), whereas direction and duration are not. The discrepancies between direction and duration have been observed (but not commented on) in previous studies employing MultiMatch (Gurtner, Bischof, and Mast, 2019; Foulsham et al., 2012; Foerster and Schneider, 2013; Li and Chen, 2018). The duration feature, which measures similarity in length of fixations throughout scanpaths, was omitted from several studies due to its degree of difference from the rest of the features (Foerster and Schneider, 2013). Further investigation into specific features and how they relate to explicit movement patterns is required in order to provide more detailed interpretation using the MultiMatch scanpath similarity analysis tool.

6.3.3 Understanding Scanpath Similarity Tool Metrics

In this thesis, we provided a simulation exploring the implications of proportions of viewing strategies on across and within conditions. This simulation provided guidelines for interpretating our scanpath results. However, as recognized in Chapter 5, our simulation only included the presence of two strategies. Further extensions of the simulation are needed that include additional strategies. Also, it may be informative to include in the simulation instances where one strategy obtains lower similarity scores overall compared to other strategies present.

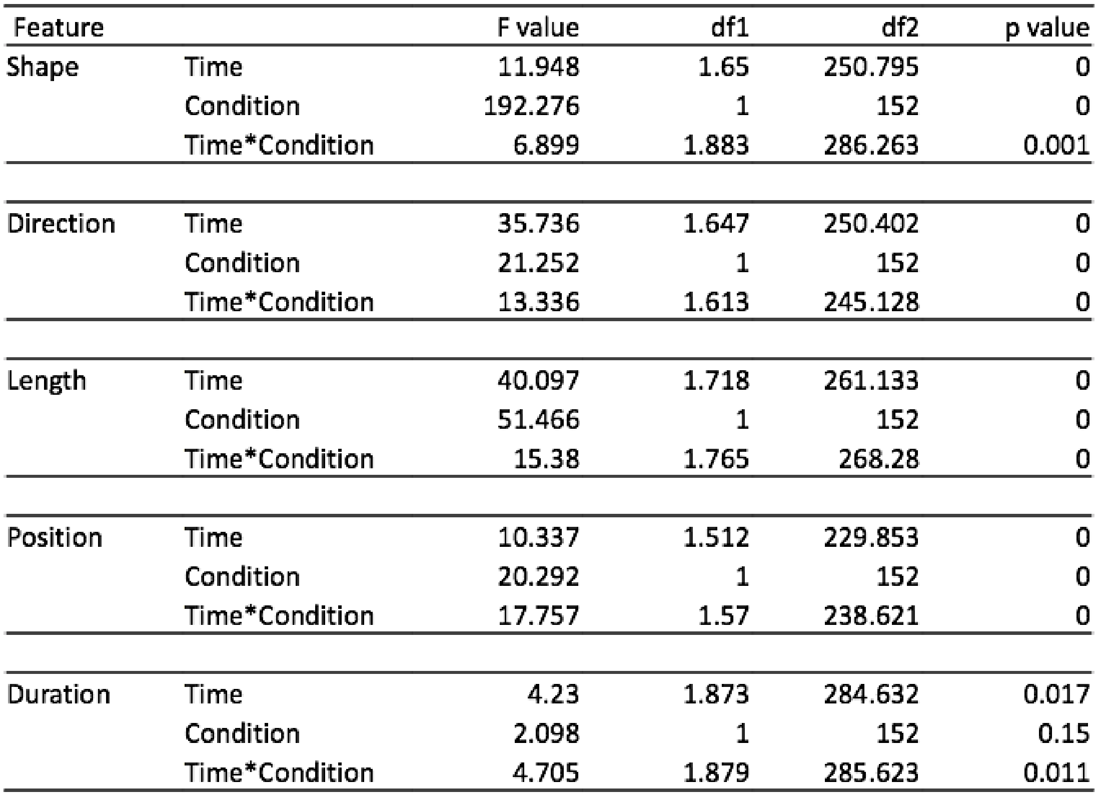
In general, additional guidelines for specific scanpath similarity analysis methods (across and within condition comparison groups) are needed to facilitate interpretation of scanpath analyses. Some studies have included randomly generated scanpaths to their analysis. These random scanpaths serve as a baseline for experimental similarity scores (Dewhurst et al., 2012; Dewhurst et al., 2018). This is something we plan to incorporate in the future to provide a baseline.

Yet another limitation of not only the present work but the scanpath field as a whole pertains to the nature of quantitative measures of scanpath similarity. Specifically, scanpath analysis on its own does not provide insight into what the visual patterns look like. Consequently, scanpath similarity does not eliminate the black box effect often encountered when examining cognitive processes.

However, scanpath similarity measures can be used in addition to other visual analysis methods to offer complementary information. For instance, prior work has supplemented scanpath analysis with extraction of the ‘most representative scanpath’ (Zhou et al., 2016), which can shed light on what a typical scanpath looks like, in turn providing insight into the corresponding hypothesized cognitive processes. The challenge, however, is that for more complex phenomena like our Analysis 2 data, this scanpath may be difficult to interpret due to its visual complexity. Thus, this is another avenue for future work.

7 Appendix A

Figure 7.1: ANOVA Table of Descriptives (F value, df1, df2, p value) for Similarity Scores over time (Problems 1, 5 and 9)



Feature		F value	df1	df2	p value
Shape	Time	11.948	1.65	250.795	0
	Condition	192.276	1	152	0
	Time*Condition	6.899	1.883	286.263	0.001
Direction	Time	35.736	1.647	250.402	0
	Condition	21.252	1	152	0
	Time*Condition	13.336	1.613	245.128	0
Length	Time	40.097	1.718	261.133	0
	Condition	51.466	1	152	0
	Time*Condition	15.38	1.765	268.28	0
Position	Time	10.337	1.512	229.853	0
	Condition	20.292	1	152	0
	Time*Condition	17.757	1.57	238.621	0
Duration	Time	4.23	1.873	284.632	0.017
	Condition	2.098	1	152	0.15
	Time*Condition	4.705	1.879	285.623	0.011

Bibliography

Ashcraft, M. H. (1992). “Cognitive arithmetic: A review of data and theory”. In: *Cognition* 44.1-2, pp. 75–106.

Ballard, D. H., M. M. Hayhoe, and J. B. Pelz (1995). “Memory representations in natural tasks”. In: *Journal of cognitive neuroscience* 7.1, pp. 66–80.

Borracci, G. et al. (2020). “The effect of assistance on learning and affect in an algebra tutor”. In: *Journal of Educational Computing Research* 57.8, pp. 2032–2052.

Bosse, M.-L. et al. (2014). “Does visual attention span relate to eye movements during reading and copying?” In: *International Journal of Behavioral Development* 38.1, pp. 81–85.

Brandt, S. A. and L. W. Stark (1997). “Spontaneous eye movements during visual imagery reflect the content of the visual scene”. In: *Journal of cognitive neuroscience* 9.1, pp. 27–38.

Bruce, N. and J. Tsotsos (2005). “Saliency based on information maximization”. In: *Advances in neural information processing systems*, pp. 155–162.

Bulling, A. et al. (2010). “Eye movement analysis for activity recognition using electrooculography”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.4, pp. 741–753.

Campbell, J. I. and M. Oliphant (1992). “Representation and retrieval of arithmetic facts: A network-interference model and simulation”. In: *Advances in psychology*. Vol. 91. Elsevier, pp. 331–364.

Castner, N. et al. (2018). “Scanpath comparison in medical image reading skills of dental students: distinguishing stages of expertise development”. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9.

Chi, M. T. et al. (1989). “Self-explanations: How students study and use examples in learning to solve problems”. In: *Cognitive science* 13.2, pp. 145–182.

Chuk, T., A. B. Chan, and J. H. Hsiao (2014). “Understanding eye movements in face recognition using hidden Markov models”. In: *Journal of vision* 14.11, pp. 8–8.

Coutrot, A. and N. Guyader (2014). “How saliency, faces, and sound influence gaze in dynamic social scenes”. In: *Journal of vision* 14.8, pp. 5–5.

Coutrot, A., J. H. Hsiao, and A. B. Chan (2018). “Scanpath modeling and classification with hidden Markov models”. In: *Behavior research methods* 50.1, pp. 362–379.

Cristino, F. et al. (2010). “ScanMatch: A novel method for comparing fixation sequences”. In: *Behavior research methods* 42.3, pp. 692–700.

Crowe, E. M., I. D. Gilchrist, and C. Kent (2018). “New approaches to the analysis of eye movement behaviour across expertise while viewing brain MRIs”. In: *Cognitive research: principles and implications* 3.1, pp. 1–14.

Dewhurst, R. et al. (2012). “It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach”. In: *Behavior research methods* 44.4, pp. 1079–1100.

Dewhurst, R. et al. (2018). “How task demands influence scanpath similarity in a sequential number-search task”. In: *Vision research* 149, pp. 9–23.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Foerster, R. M. and W. X. Schneider (2013). “Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units". In: *Journal of Eye Movement Research* 6.5.

Foulsham, T. and G. Underwood (2008). "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition". In: *Journal of vision* 8.2, pp. 6–6.

Foulsham, T. et al. (2012). "Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach". In: *Journal of Eye Movement Research* 5.4, pp. 1–14.

French, R. M., Y. Glady, and J.-P. Thibaut (2017). "An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making". In: *Behavior research methods* 49.4, pp. 1291–1302.

Groen, G. J. and J. M. Parkman (1972). "A chronometric analysis of simple addition." In: *Psychological review* 79.4, p. 329.

Gurtner, L. M., W. F. Bischof, and F. W. Mast (2019). "Recurrence quantification analysis of eye movements during mental imagery". In: *Journal of vision* 19.1, pp. 17–17.

Hacisalihzade, S. S., L. W. Stark, and J. S. Allen (1992). "Visual perception and sequences of eye movement fixations: A stochastic modeling approach". In: *IEEE Transactions on systems, man, and cybernetics* 22.3, pp. 474–481.

Haselen, G. L.-V., J Van Der Steen, and M. Frens (2000). "Copying strategies for patterns by children and adults". In: *Perceptual and Motor Skills* 91.2, pp. 603–615.

Henderson, J. M. and F. Ferreira (2004). "Scene perception for psycholinguists." In:

Holmqvist, K. et al. (2011). "A method for quantifying focused versus overview behavior in AOI sequences". In: *Behavior research methods* 43.4, pp. 987–998.

Huebner, M. G. and J.-A. LeFevre (2018). “Selection of procedures in mental subtraction: Use of eye movements as a window on arithmetic processing.” In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 72.3, p. 171.

Itti, L. and C. Koch (2000). “A saliency-based search mechanism for overt and covert shifts of visual attention”. In: *Vision research* 40.10-12, pp. 1489–1506.

Jarodzka, H. et al. (2010). “In the eyes of the beholder: How experts and novices interpret dynamic stimuli”. In: *Learning and Instruction* 20.2, pp. 146–154.

Jarodzka, H. et al. (2012). “Conveying clinical reasoning based on visual observation via eye-movement modelling examples”. In: *Instructional Science* 40.5, pp. 813–827.

Jennings, J. and K. Muldner (2020). “Assistance that fades in improves learning better than assistance that fades out”. In: *Instructional Science* 48.4, pp. 371–394.

Jost, T. et al. (2004). “Contribution of depth to visual attention: comparison of a computer model and human”. In: *Proceedings. Early cognitive vision workshop*. 3D Modelling, pp. 1–4.

Jovancevic, J., B. Sullivan, and M. Hayhoe (2006). “Control of attention and gaze in complex environments”. In: *Journal of Vision* 6.12, pp. 9–9.

Judd, T. et al. (2009). “Learning to predict where humans look”. In: *2009 IEEE 12th international conference on computer vision*. IEEE, pp. 2106–2113.

Koehler, M. J. et al. (2014). “The technological pedagogical content knowledge framework”. In: *Handbook of research on educational communications and technology*. Springer, pp. 101–111.

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

Lee, H. S., S. Betts, and J. R. Anderson (2015). “Not taking the easy road: When similarity hurts learning”. In: *Memory & cognition* 43.6, pp. 939–952.

LeFevre, J.-A. et al. (1996). “Multiple routes to solution of single-digit multiplication problems.” In: *Journal of Experimental Psychology: General* 125.3, p. 284.

Levenshtein, V. I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union, pp. 707–710.

Li, A. and Z. Chen (2018). “Representative Scanpath Identification for Group Viewing Pattern Analysis”. In: *Journal of Eye Movement Research* 11.6.

Mannan, S, K. H. Ruddock, and D. S. Wooding (1995). “Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images.” In: *Spatial vision*.

Mathôt, S. et al. (2012). “A simple way to estimate similarity between pairs of eye movement sequences”. In: *Journal of Eye Movement Research* 5.1.

Mauro, D. G., J.-A. LeFevre, and J. Morris (2003). “Effects of problem format on division and multiplication performance: Division facts are mediated via multiplication-based representations.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.2, p. 163.

Mayer, R. E. (2010). “Unique contributions of eye-tracking research to the study of learning with graphics”. In: *Learning and instruction* 20.2, pp. 167–171.

McIntyre, N. A. and T. Foulsham (2018). “Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms”. In: *Instructional Science* 46.3, pp. 435–455.

Muldner, K. and C. Conati (2010). “Scaffolding meta-cognitive skills for effective analogical problem solving via tailored example selection”. In: *International Journal of Artificial Intelligence in Education* 20.2, pp. 99–136.

Nivala, M. et al. (2016). “Developing visual expertise in software engineering: An eye tracking study”. In: *2016 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, pp. 613–620.

Noton, D. and L. Stark (1971). “Scanpaths in saccadic eye movements while viewing and recognizing patterns”. In: *Vision research* 11.9, 929–IN8.

Posner, M. I. and Y. Cohen (1984). “Components of visual orienting”. In: *Attention and performance X: Control of language processes* 32, pp. 531–556.

Posner, M. I. et al. (1985). “Inhibition of return: Neural basis and function”. In: *Cognitive neuropsychology* 2.3, pp. 211–228.

Rajashekar, U. et al. (2008). “GAFFE: A gaze-attentive fixation finding engine”. In: *IEEE transactions on image processing* 17.4, pp. 564–573.

Reed, S. K. (2012). “Learning by mapping across situations”. In: *Journal of the Learning Sciences* 21.3, pp. 353–398.

Reed, S. K. and C. A. Bolstad (1991). “Use of examples and procedures in problem solving.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17.4, p. 753.

Reed, S. K., A. Dempster, and M. Ettinger (1985). “Usefulness of analogous solutions for solving algebra word problems.” In: *Journal of experimental psychology: Learning, Memory, and Cognition* 11.1, p. 106.

Reingold, E. M. et al. (2001). “Visual span in expert chess players: Evidence from eye movements”. In: *Psychological science* 12.1, pp. 48–55.

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

Siegler, R. S. (1989). “Strategy diversity and cognitive assessment”. In: *Educational researcher* 18.9, pp. 15–20.

Stirk, J. A. and G. Underwood (2007). "Low-level visual saliency does not predict change detection in natural scenes". In: *Journal of vision* 7.10, pp. 3–3.

Stranc, S. and K. Muldner (2020). "Scanpath Analysis of Student Attention During Problem Solving with Worked Examples". In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 306–311.

Stranc, S., S. Tan, and K. Muldner (2020). "Scanpaths distinguish problem format in a math cognition task". In:

Susac, A. N. et al. (2014). "EYE MOVEMENTS REVEAL STUDENTS'STRATEGIES IN SIMPLE EQUATION SOLVING". In: *International Journal of Science and Mathematics Education* 12.3, pp. 555–577.

Sweller, J. and G. A. Cooper (1985). "The use of worked examples as a substitute for problem solving in learning algebra". In: *Cognition and instruction* 2.1, pp. 59–89.

Tan, S., K. Muldner, and J.-A. LeFevre (2016). "Solution of division by access to multiplication: Evidence from eye tracking." In: *CogSci*.

Torgerson, W. S. (1952). "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4, pp. 401–419.

Tricot, A. and J. Sweller (2014). "Domain-specific knowledge and why teaching generic skills does not work". In: *Educational psychology review* 26.2, pp. 265–283.

Turano, K. A., D. R. Geruschat, and F. H. Baker (2003). "Oculomotor strategies for the direction of gaze tested with a real-world activity". In: *Vision research* 43.3, pp. 333–346.

VanLehn, K. (1996). "Cognitive skill acquisition". In: *Annual review of psychology* 47.1, pp. 513–539.

VanLehn, K. (1998). "Analogy events: How examples are used during problem solving". In: *Cognitive Science* 22.3, pp. 347–388.

VanLehn, K. (1999). "Rule-learning events in the acquisition of a complex skill: An evaluation of CASCADE". In: *The Journal of the Learning Sciences* 8.1, pp. 71–125.

Wang, W. et al. (2011). "Recent advances in catalytic hydrogenation of carbon dioxide". In: *Chemical Society Reviews* 40.7, pp. 3703–3727.

Wang, Y. et al. (2017). "Scanpath estimation based on foveated image saliency". In: *Cognitive processing* 18.1, pp. 87–95.

Widaman, K. F. et al. (1989). "A componential model for mental addition." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15.5, p. 898.

Widaman, K. F. et al. (1992). "Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models". In: *Learning and Individual Differences* 4.3, pp. 167–213.

Zhang, H., N. C. Anderson, and K. F. Miller (2021). "Refixation patterns of mind-wandering during real-world scene perception." In: *Journal of experimental psychology: human perception and performance* 47.1, p. 36.

Zhou, L. et al. (2016). "A scanpath analysis of the risky decision-making process". In: *Journal of Behavioral Decision Making* 29.2-3, pp. 169–182.