Behav Res
DOI 10.3758/s13428-016-0788-z

An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making

Robert M. French[1] · Yannick Glady[1] · Jean-Pierre Thibaut[1]

Abstract In recent years, eyetracking has begun to be used to study the dynamics of analogy making. Numerous scanpath-comparison algorithms and machine-learning techniques are available that can be applied to the raw eyetracking data. We show how scanpath-comparison algorithms, combined with multidimensional scaling and a classification algorithm, can be used to resolve an outstanding question in analogy making—namely, whether or not children's and adults' strategies in solving analogy problems are different. (They are.) We show which of these scanpath-comparison algorithms is best suited to the kinds of analogy problems that have formed the basis of much analogy-making research over the years. Furthermore, we use machine-learning classification algorithms to examine the item-to-item saccade vectors making up these scanpaths. We show which of these algorithms best predicts, from very early on in a trial, on the basis of the frequency of various item-to-item saccades, whether a child or an adult is doing the problem. This type of analysis can also be used to predict, on the basis of the item-to-item saccade dynamics in the first third of a trial, whether or not a problem will be solved correctly.

Keywords Eyetracking algorithms · Jarodzka algorithm · LDA · SVM · Analogy strategies

Traditionally, analogy making has been studied statically. Participants typically see a pair of related images (the "base pair"), along with a third image and a number of candidate target images. One of these target images—the "correct analogical match"—is supposed to be related to the third image in the same way that the base items were related to one another. The participant's task is to identify the correct analogical match. Correct and incorrect answers (and, sometimes, reaction times) are recorded and analyzed. However, these studies could not capture—and in fairness, were not designed to capture—the *dynamic* aspects of solving an analogy problem. As such, they shed essentially no light on the question of what strategies were adopted during the course of solving analogy problems.

In this article, we introduce a novel means of studying the dynamic aspects of analogy making in both children and adults. The proposed methodology involves combining eyetracking, multidimensional scaling (MDS), and neural-network classification algorithms, as well as using machine-learning algorithms to analyze the component vectors making up participants' scanpaths. In what follows, we will briefly describe each of these techniques and show how they can be combined successfully in the context of analogy making.

Although the purpose of this article is, first and foremost, a methodological one, it is important to note that the development of these techniques has allowed us (French & Thibaut, 2014; Thibaut & French, 2016; Thibaut, French, Missault, Gérard, & Glady, 2011) to answer, for what we believe to be the first time, a long-standing question in the field of analogy-making—namely, do children and adults use the same (or very similar) search-space strategies when solving analogy problems? The answer, as will be shown in what follows, is "no."

Eyetracking

Eyetracking involves following the gaze trajectories of participants as they perform a particular task. The underlying

✉ Robert M. French
robert.french@u-bourgogne.fr

[1] LEAD-CNRS UMR 5022, Université de Bourgogne-Franche-Comté, Pôle AAFE, 11 Esplanade Erasme, 21000 Dijon, France

Published online: 09 September 2016

assumption is that sequences of eye movements (i.e., scanpaths) are a reflection of the mental activity involved in studying a scene, examining a face, pondering a configuration of items, and so on. It is the first tool that has allowed the dynamics of solving analogy problems to be studied.

Analyzing eyetracking data

Obviously, recording participants' scanpaths as they do analogy problems is of little use unless these data can be analyzed in an appropriate manner. There are currently a number of different scanpath-comparison techniques, each with its own advantages and disadvantages. In the present article, we compare three of the most important of these techniques in the context of their application to the study of analogy making. To compare these techniques, we analyze their outputs by means of multidimensional scaling and neural-network classification algorithms.

The test bed for these techniques will be how well these algorithms can be used to answer what for many years has been an open question in the field of analogy making—namely, whether or not children's analogy problem-solving strategies are different from those of adults. One of these techniques, developed by Jarodzka, Holmqvist, and Nyström (2010), allows us to answer this question (in the affirmative) significantly better than the other two.

Subsequently, we analyze the item-to-item gaze transitions making up these scanpaths using two different machine-learning classification algorithms, linear discriminant analysis (LDA; Fisher, 1936) and support vector machines (SVM; Vapnik, 1995, 1998). These techniques not only allow us to better understand *where* the differences between adults' and children's search strategies lay and at what point in time these differences arise, but crucially, they also allow us to *predict*, significantly better than chance and very early in a trial, whether a child or an adult was doing the problem, whether or not the problem would be solved correctly, and so forth.

Background

Analogical reasoning is a ubiquitous process in thinking and reasoning (Gentner & Smith, 2012; Hofstadter, 2001; Holyoak, 2012; Holyoak, Gentner, & Kokinov, 2001). It can be defined as a comparison of two domains (the source and the target domains) on the basis of their respective relational structure (Gentner, 1983). Studies of analogy making have explored two main explanations for its development—namely, the increase of structured knowledge (Gentner & Rattermann, 1991; Goswami, 1992; Goswami and Brown, 1990) and the maturation of executive functions (Halford, 1993; Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2010a, 2010b). An important prediction of the executive-function view is that children and adults should organize their searches of the analogy-problem space differently (see also Woods et al., 2013). This is what we mean when we say that they use different strategies when solving analogy problems. What information is sought and how the search for this information is organized in time are crucial to understanding how the analogy problem is solved. Attention and gaze fixations are highly correlated, especially for complex stimuli (Deubel & Schneider, 1996; He & Kowler, 1992), and the fixation time for a given object is correlated with its informativeness in a scene (Nodine, Carmody, & Kundel, 1978). In other words, eye movements can provide a window on specific problem-solving strategies—in particular, for problems involving visual information. This makes eyetracking particularly well adapted to the types of analogy problems that we will consider.

We are not the first to use eyetracking technology to study analogy making, but this type of analysis remains, nonetheless, in its infancy. Eyetracking techniques were first used by Bethell-Fox, Lohman, and Snow (1984) to study strategies when reasoning by analogy. They found strategic differences in adults with high or low fluid intelligence when solving geometric A:B::C:? problems. More recently, Gordon and Moser (2007) investigated adults' strategies in scene analogy problems. Thibaut et al. (2011), Glady et al. (2013), French and Thibaut (2014), and Thibaut and French (2016) all used eyetracking technology to examine children's gaze locations and item-to-item transitions during analogy tasks, demonstrating clear differences between adults' and children's strategies in solving analogy problems.

Comparing three scanpath-comparison algorithms

A scanpath is the complete visual trajectory of a participant's eye movements during a task, and various techniques have been developed to characterize and compare scanpaths. We will consider three of these techniques: The most widely used is an algorithm developed by Levenshtein (1966), another is the widely used attentional map algorithm (AMAP; Ouerhani, von Wartburg, Hugli, & Muri, 2004; Rajashekar, Van der Linde, Bovik, & Cormack, 2008), and the third is a relatively recent vector-based algorithm developed by Jarodzka, Holmqvist, and Nyström (2010). Each of these algorithms compares two scanpaths and produces a number that indicates how similar they are to each other. We will compare these three scanpath algorithms according to how well they distinguish children's from adults' scanpaths during analogy-solving problems. All three of these scanpath algorithms showed that there were, in fact, significant differences between how children and adults solve analogy problems. However, one of the algorithms, the Jarodzka et al. algorithm, is best suited to these analyses and outperforms the other two.

Scanpath comparison

To do this comparison, we gave children and adults the same analogy problems and recorded their scanpaths while they were solving the problems. We then used each of the scanpath-comparison algorithms to produce a pairwise comparison of all of the scanpaths, both children's and adults', to produce a similarity matrix between all scanpaths for the problems. By means of multidimensional scaling (MDS; Cox & Cox, 2001; Torgerson, 1952) we converted this matrix into a 2-D map that reflected each of the similarity measures. Each scanpath was represented by a point on this 2-D MDS map (see Fig. 4a–c below). We then performed a "leave-one-out cross-validation" (LOOCV) procedure (see Geisser, 1975; Lachenbruch, 1967; Miller, 1974; Stone, 1974; for a review, see Arlot & Celisse, 2010) on these points using a standard feedforward–backpropagation (FFBP) network (Rumelhart, McClelland, & the PDP Research Group, 1986). This analysis worked as follows. For each point *p* in the MDS map, we trained the FFBP network to correctly classify (i.e., adult or child) all of the other points in the map except *p* (hence, the name of the procedure, "leave-one-out"). We then presented the previously unseen point, *p*, to the network to see whether the network would classify *p* correctly (i.e., as to whether it corresponded to an adult's or a child's scanpath). We did this for all points *p* in the 2-D MDS map. For all of the scanpath-comparison algorithms, once the dimensionality of the data had been reduced by MDS, the FFBP network was able to correctly classify the left-out participants well above chance, which shows that adults and children do use different strategies to solve analogy problems. As we will show in more detail below, the Jarodzka et al. (2010) algorithm produced the best results.

We begin with a brief description of each of the three algorithms we tested.

Levenshtein's (1966) "string-edit" algorithm

This algorithm divides the scan area into predefined areas of interest (AOIs) and then associates each of the fixation coordinates recorded by the eyetracker with one of these areas. Scanpaths are considered to be sequences of these AOIs. The duration of the fixation in each area is not taken into account (i.e., consecutive fixations that fall into one AOI are collapsed). Suppose, for example, that the AOIs for a particular problem are labeled A, B, C, D, E, F, G, and H. Suppose further that there is a scanpath S_1 = BADEGAGCB, which meant that the participant's gaze moved successively from area B to A to D to E, and so forth. A second, shorter scanpath might be S_2 = ABDEGBG. The Levenshtein algorithm is a "string-edit" algorithm that determines the "distance" between two scanpaths as the smallest number of single-letter substitutions, deletions, and/or insertions required to transform one string into the other. This number is calculated using the Wagner–Fischer algorithm (Wagner & Fischer, 1974) and is the *Levenshtein distance* between the two scanpaths.

Attention map (AMAP) scanpath comparison

There are a number of "attention map" algorithms. These algorithms compare two scanpaths by computing how long various locations are looked at, how far each fixation point in one scanpath is from the closest fixation point in the other scanpath, and so forth. One of the earliest algorithms based on attention measures is the Mannan distance algorithm (Mannan, Ruddock, & Wooding, 1997). However, this class of scanpath-comparison techniques has a number of drawbacks—most importantly for our purposes, the temporal order of fixations is lost. So, even if the two scanpaths have very different lengths and shapes, an AMAP algorithm could still indicate a high degree of similarity between them (Le Meur & Baccino, 2013). When attempting to uncover exploration strategies that unfold over time, the loss of temporal information poses a serious problem. More recent AMAP comparison algorithms (e.g., Ouerhani et al., 2004; Rajashekar et al., 2008) create attention "landscapes" by accumulating fixed-width Gaussians over fixation points. It is generally accepted that the longer a fixation time on a particular item, the deeper the visual processing of that item (Just & Carpenter, 1976). In this attentional-landscape algorithm, as in the earliest AMAP algorithms, temporal-order information is still lost.

After obtaining attention maps for each trial, comparison scores between the different scanpaths are obtained using a coefficient of correlation between the values of the two attention maps. As with the Levenshtein algorithm, we used the AMAP pairwise scanpath-comparison scores to create a similarity matrix comparing children's and adults' scanpaths for the three sets of problems described above.

Vector-based scanpath-comparison (Jarodzka et al., 2010)

A novel method of scanpath comparison was recently proposed by Jarodzka et al. (2010). This algorithm turns out to be a particularly powerful one for analyzing scanpaths from analogy-making problems. Below we present our simplification of this algorithm.

A scanpath is considered to be made up of a series of "saccade vectors,"—that is, a concatenated series of vectors whose endpoints correspond to the coordinates of successive gaze points (Fig. 1, left panel). The scanpath is first simplified by combining into a single vector any two consecutive saccade vectors that are nearly collinear and by combining very short vectors with longer adjacent ones (Fig. 1, right panel). In general, very small saccade vectors occur when a participant has fixed his or her gaze on a particular item.

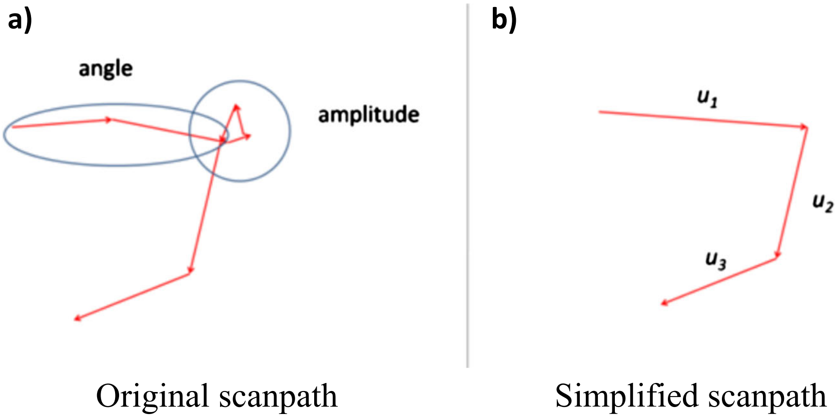


Fig. 1 Simplifying scanpaths according to the Jarodzka et al. (2010) algorithm

After this simplification, two scanpath vectors can be compared by "stretching" one or both of them appropriately. *Scanpath stretching*, which is at the heart of this algorithm, requires some explaining. Assume that there are two saccade vectors, U = {u_1, u_2, u_3} and V = {v_1, v_2, v_3, v_4}. In other words, scanpath U consists of the saccade vector u_1 followed by u_2, which is followed by u_3. Similarly, the scanpath V consists of saccade vector v_1, followed by v_2, followed by v_3, followed by v_4. To compare U and V, we need to transform them into two scanpaths of the same length. To achieve this, we "stretch" the scanpaths, as necessary, so that we can align them for comparison. This is done by adding immediate repetitions of saccade vectors (we call this "stretching" the original scanpath), so that the two stretched scanpaths have the same length. Our goal is to find the two stretched scanpaths, U' and V', that are as similar as possible to each other with respect to the chosen similarity metric (orientation, length, etc.). The degree of similarity between U' and V' will be the measure of the similarity between U and V.

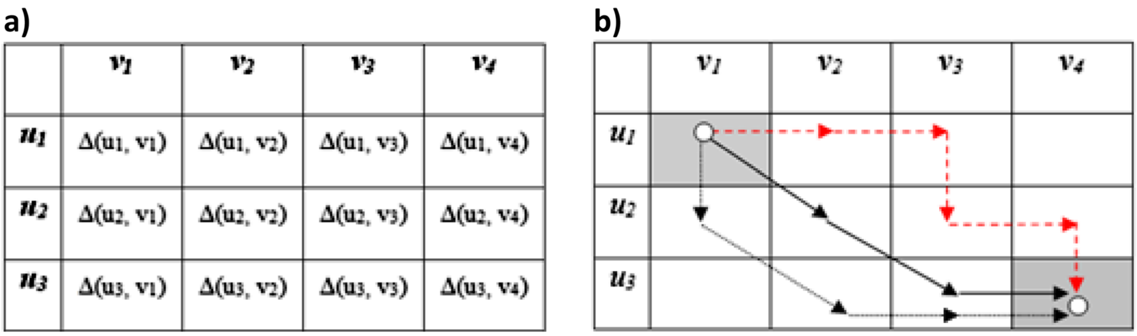
The idea is to make a matrix with the saccade vectors of one scanpath on the *x*-axis and the saccade vectors of the second scanpath on the *y*-axis (see Fig. 2). The uppermost cell on the left is the starting cell, and the lowermost cell on the right is the ending cell. We then traverse this matrix from the starting cell to the ending cell, on each step always moving closer to the ending cell. ("Backward" moves are not permitted.) Each cell that is traversed contains a value that measures how close the two saccade vectors associated with that cell are. (The lower the value, the more similar the two saccade vectors.) Our goal is find the path with the lowest possible total similarity value.

So, if we suppose that the path through the matrix that goes through {(u_1, v_1), (u_1, v_2), (u_1, v_3), (u_2, v_3), (u_2, v_4), (u_3, v_4)} (shown in dashed red in Fig. 2b) is the one with the smallest total difference value, we observe that U has been "stretched" to become U' by repeating u_1 and u_2 to become U' = {u_1, u_1, u_1, u_2, u_2, u_3}, and V has been stretched by repeating v_3 and v_4 to become V' = {v_1, v_2, v_3, v_3, v_4, v_4}.

U' and V' now have the same length and can, therefore, be compared by a pairwise comparison of their respective component saccade vectors. This comparison may be made on the basis of the respective lengths of the paired component saccade vectors, their orientations, and so forth.

We now describe this algorithm in detail. A saccade vector difference matrix is first created (Fig. 2a). Each of the saccade vectors making up one of the scanpaths is compared to each of the saccade vectors making up the other scanpath, according to a metric—generally, vector magnitude or orientation (magnitude, in our study). Once this table is constructed, we consider all paths through the table that begin with the comparison of the first saccade vectors in both scanpaths [i.e., cell (1, 1) of the table, containing $\Delta(u_1, v_1)$] and end with a comparison of the final saccade vectors in each scanpath [i.e., cell (3, 4) of the table, containing $\Delta(u_3, v_4)$]. The traverse of the difference matrix always moves to the right, down, or diagonally down-and-right. Three examples of paths through the matrix are illustrated in the Fig. 2b. Each path through the table corresponds to the comparison of two specific (stretched) scanpaths. For example, the uppermost path shown corresponds to a comparison between U' = {u_1, u_1, u_1, u_2, u_2, u_3} and V' = {v_1, v_2, v_3, v_3, v_4, v_4}. This path corresponds to the sum of the values in the cells (1, 1), (1, 2), (1, 3), (2, 3), (2, 4), (3, 4) of the saccade vector difference matrix. When all of these paths through the matrix are considered, the path that has the smallest total difference value (i.e., the smallest cumulative sum of comparisons) is selected. This path corresponds to the two stretched scanpaths that are the most similar.

We simplified the Jarodzka et al. (2010) algorithm by eliminating the relatively complex Dijkstra (1959) tree-search algorithm that it uses. Instead, we simply constructed a path through the difference matrix by moving only rightward, downward, or diagonally from the upper-left cell toward the lower-right cell. As we progressed incrementally through the saccade vector difference matrix, we recorded in the cells of the cumulative-difference matrix in Fig. 2b the smallest sum of the difference values



a)

	v_1	v_2	v_3	v_4
u_1	$\Delta(u_1, v_1)$	$\Delta(u_1, v_2)$	$\Delta(u_1, v_3)$	$\Delta(u_1, v_4)$
u_2	$\Delta(u_2, v_1)$	$\Delta(u_2, v_2)$	$\Delta(u_2, v_3)$	$\Delta(u_2, v_4)$
u_3	$\Delta(u_3, v_1)$	$\Delta(u_3, v_2)$	$\Delta(u_3, v_3)$	$\Delta(u_3, v_4)$

Fig. 2 (**a**) Saccade vector difference matrix. Each of the saccade vectors making up each of the two scanpaths is compared on the basis of the chosen metric, and a saccade vector difference table is drawn up containing these differences. (**b**) Cumulative-difference matrix. The comparison of each pair of stretched scanpaths corresponds to a traverse of the table from the upper-left to the lower-right corner of the saccade vector difference matrix (the only directions of movement permitted are down, right, and diagonally down-and-right). We find the path that produces the lowest total difference value, and this value is the measure of the similarity between U and V

of all the paths that led to that cell. This is similar to the matrix-traversal technique used in the Wagner–Fischer algorithm (Wagner & Fischer, 1974), adopted in the Levenshtein string-edit algorithm. Necessarily, more than one path leads to most cells (except cells at the top and left edges of the matrix). Thus, in each cell, we put the value of the "least costly" path to that cell, which was the path corresponding to the greatest overall similarity of the scanpaths to that point. This meant that at each step of the process, each cell of the cumulative-difference matrix always contained the value of the "least costly" path from C(1, 1) to that cell. The difference measure between any two scanpaths U and V is the cumulative sum of the differences in the lower-right cell of the cumulative-difference matrix, normalized by the number of steps taken through the matrix.

As we had done for the Levenshtein and AMAP algorithms, we used the Jarodzka et al. (2010) algorithm to create a similarity matrix between the adults' and children's scanpaths for the four trials in each of the three conditions (see the Materials section in the description of the experiment below, as well as Fig. 3). The metric we used for the similarity of the saccade vectors (i.e., to calculate the saccade vector difference matrix for each pair of scanpaths) was their length. Using a standard MDS (Torgerson, 1952) procedure, we transformed the similarity matrices into 2-D scatterplots (see Fig. 4 in the Results below).

Testing the scanpath algorithms and analyzing their component item-to-item transitions

To test the performances of the three scanpath-comparison algorithms described above in the domain of analogy making, and to examine further the information that can be gleaned from item-to-item transitions within these scanpaths, we ran an analogy-making experiment composed of three different types of analogy-making tasks.

Experiment with three analogy-making tasks

Overview

The goal of this experiment was to consider the output of each of the three scanpath-comparison algorithms for a set of three different types of analogy problems done by children and adults. These data were then converted by MDS into a 2-D plot and analyzed by means of a neural-net classifier to determine how well each of the scanpath algorithms discriminated children's scanpaths from those of adults.

Method

Participants The participants were 20 adults (14 females, six males; mean age = 20.4 years, *SD* = 2.21; range: 17 to 27), who were students at the University of Burgundy–Franche-Comté and naïve to analogical reasoning tasks, and 25 six-year-olds (16 females, nine males; mean age = 79.5 months, *SD* = 3.6; range: 73 to 84). For the children participating in this experiment, the parents' informed consent was obtained.

Materials Three tasks, each composed of three training trials and four experimental trials, constituted the experiment. The first task was a "Scene" analogy problem task (Richland et al., 2006), the second a standard A:B::C:? task (called "ABCD"), and the third an A:B::C:? task with the items composing the problems put into a context (e.g., a bird flying to its nest, etc.; hereafter called "ABCD-scene"). Each problem of each task was composed of seven images, each being a black-and-white line drawing (Fig. 3).

In the scene analogy problems ("Scene"), the top scene was composed of two elements depicting a binary semantic relation: in Fig. 3, a mouse (A) being chased by a cat (B). One of these two elements (B) had an arrow pointing to it. The bottom scene was composed of five drawings: the two elements

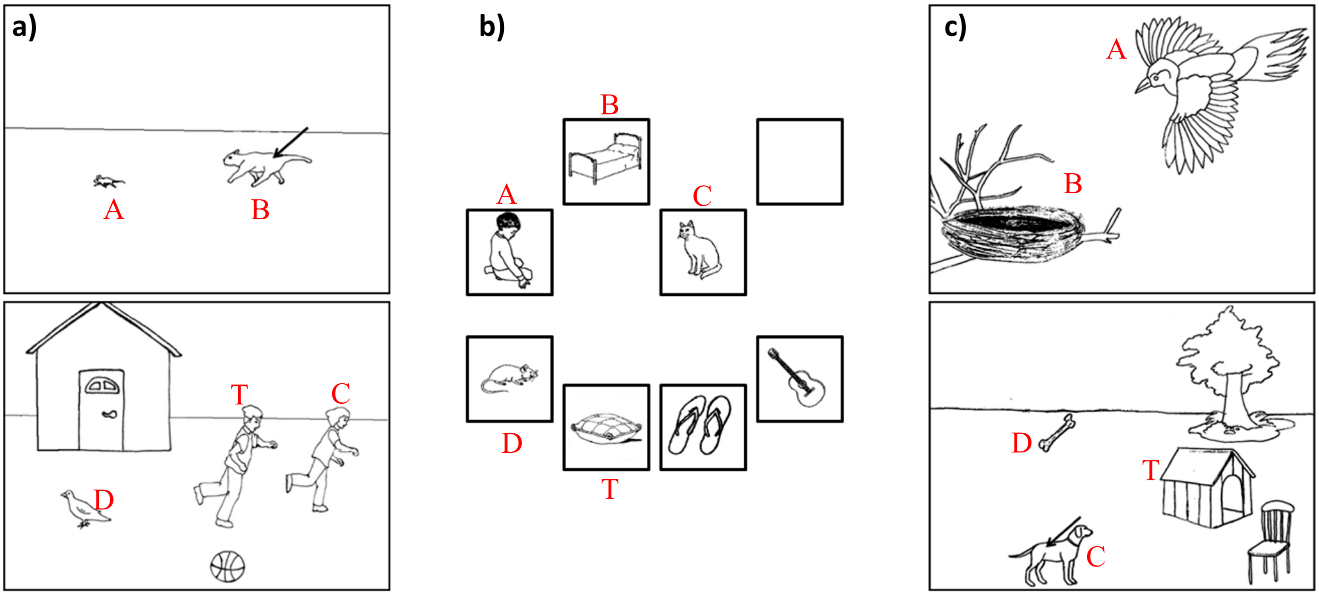


Fig. 3 Presentations of the three tasks used for this experiment: (**a**) the scene analogy task ("Scene"), (**b**) a standard A:B::C:? task ("ABCD"), and (**c**) the scene-oriented A:B::C:? task ("ABCD-Scene")

depicting the same relation as in the top picture: here, a girl (C) being chased by a boy (T). There was also a distractor item, in this case a bird (D), and two elements that were consistent with the scene but that had no salient relation with the elements of the relation. These pictures (501 × 376 pixels) were based on those created by Richland et al. (2006). We have labeled the items in the scene analogy problem to correspond to the A:B::C:D paradigm.

In the standard A:B::C:? task ("ABCD"), the A, B, and C drawings were presented in the top row, along with an empty square symbolizing the location of the solution. The four remaining pictures, the target (T), a related-to-C distractor (D), and two unrelated distractors, were presented in a row at the bottom of the screen. The size of each picture was 200 × 195 pixels.

The contextualized A:B::C:? task ("ABCD-Scene") consisted of two scenes (501 × 376 pixels). The top picture was composed of two black-and-white line drawings with a relation between them. In Fig. 3, this is a bird (A) flying to its nest (B). The bottom picture was composed of five drawings: in the figure, a dog (C), a doghouse (T), a bone (the semantic distractor, D), and two unrelated distractors. This task differed from the first task in that here the C term was designated with an arrow, and not one of the elements constituting the base relation. It differed from the second task because the different pictures constituting the problem were grouped into two scenes, but it was otherwise equivalent to the standard A:B::C:? task. The materials of the last two tasks were based on those previously used by Thibaut et al. (2011).

The tasks were displayed on a Tobii T120 eyetracker device with a 1,024 × 768 screen resolution. A standard five-point calibration for the eyetracker was used. Prior to each trial, an image of a duck was presented in the middle of the screen instead of the standard fixation cross.

Procedure Appropriate controls were carried out to ensure that the participants knew what the items in each of the problems were and that they understood the instructions. In the first task, they were asked to point to the element in the bottom scene that played the same role as the one that had an arrow pointing to it in the top scene. The two others tasks were administered as in Thibaut et al. (2011). Eyetracking data were gathered from the moment of the initial presentation of the problem to the moment that a choice of one of the answers was made. The participant's scanpath for a particular problem consisted of a record of his or her gaze-fixation points, taken every 8 ms.

Analysis of the data Using the three different scanpath-comparison algorithms described above, we compared the scanpaths of adults and children on strictly identical problems. It was, of course, necessary for each problem to be seen by both adults and children, so that the locations of the items were identical. Using each of the three scanpath-comparison algorithms, we created three similarity matrices for the full set of scanpaths, one for each algorithm. These matrices, which were subsequently analyzed by an MDS algorithm, were produced by performing a pairwise comparison of all of the children's and all of the adults' scanpaths. In other words, the matrices consisted of all child–child, child–adult, and adult–adult scanpath comparisons.

Results

MDS scatterplots of children's and adults' scanpaths

Below we show the MDS scatterplots (Fig. 4) derived from the similarity matrices computed by each of the three scanpath-comparison algorithms for the trials in each of the three experimental conditions. (See the Materials section of the experiment description above and the examples shown in Fig. 3.)

Each of the points (o's and x's) in these scatterplots represents a scanpath, for either an adult (o) or a child (x), recorded as the participant solved one of the three types of analogy problems. The extent to which the points for children clumped in distinct groups that were different from those of adults was a measure of how distinct the analogy-solving strategies were for the two groups. We can see that both groups of points for the scatterplots produced by the Levenshtein algorithm are quite

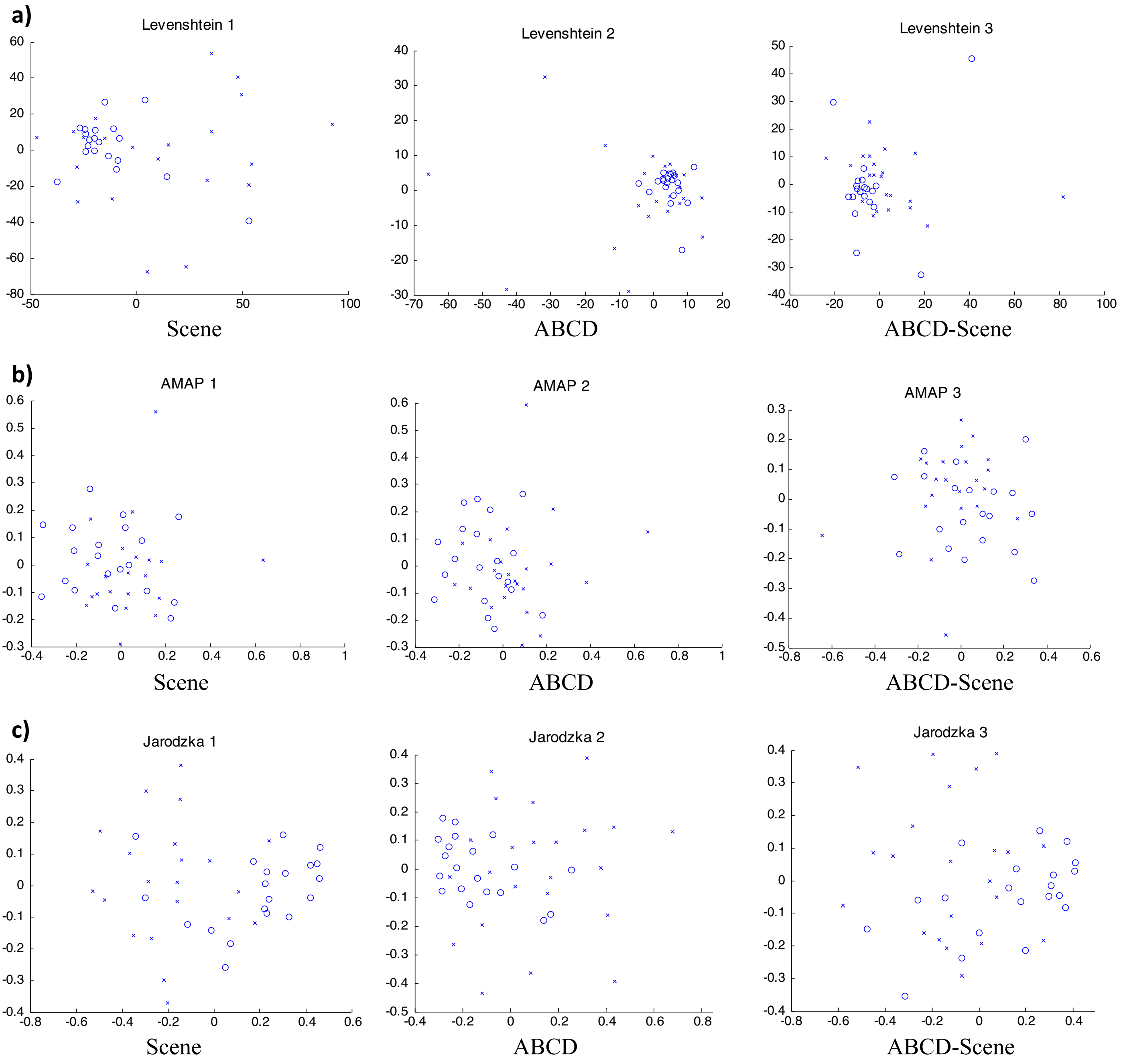


Fig. 4 (**a**) Multidimensional scaling (MDS) scatterplots derived from the scanpath similarity matrices produced by Levenshtein's (1966) algorithm. (**b**) MDS scatterplots derived from the scanpath similarity matrices produced by an attention-mapping algorithm (AMAP: Ouerhani et al., 2004; the x's are children, the o's adults). (**c**) MDS scatterplots derived from the scanpath similarity matrices produced by Jarodzka et al.'s (2010) algorithm (the x's are children, the o's adults)

tightly clustered together, that those produced by the AMAP algorithm are far more dispersed and hard to distinguish, and that those produced by the Jarodzka algorithm are the easiest to distinguish. In the next section, we quantify these differences.

Neural-network classification of the MDS scanpath scatterplot points For each of the conditions and each of the scanpath classification algorithms, we wished to quantify the extent to which the scanpaths from the adults were distinct from those of the children. To do this, we used a standard LOOCV procedure on the points in the MDS map using a standard FFBP network (Rumelhart et al., 1986). Specifically, we used a three-layer perceptron with two input units (one for each coordinate of the points in the MDS map), five hidden units, and one category node (i.e., child or adult). There was a bias node on the input and hidden layers. During training, the network was run either until all of its training exemplars had been learned to a 0.2 criterion or for a maximum of 2,500 training epochs. We used a shallow sigmoid with a temperature parameter (β) of 0.1. For each MDS map, the input to the network consisted of the real coordinates of each point in the map, and the "teacher" for that point was the group (adult/child) to which it belonged.

We ran an LOOCV procedure for all of the points in each MDS map. We then computed the total number of points that had been correctly classified. The higher this value, the more distinct were the scanpaths of adults and children.

The results of this analysis are shown in Fig. 5. All results are significantly above chance (i.e., .5). Of the three scanpath-comparison algorithms, the performance of the Jarodzka et al. (2010) algorithm (with "vector magnitude" as the comparison metric) is the best, and the AMAP algorithm the poorest. In the case of the Jarodzka et al. algorithm, we obtained an adult/child prediction accuracy of 80 % for the scene analogy problems.

Studying the item-to-item saccades (transitions) making up the scanpaths Once we had looked at the analyses of the global scanpaths, we then considered the item-to-item saccades (transitions) that made up the scanpaths. We did this on the basis of the idea that if a participant had frequent successive saccades between two items, then he or she was considering that there was some relation between those two items, a relation that was, or might be, important in solving the analogy problem. The importance of the role of the relations between individual items is almost universally accepted in the analogy-making community. We believe that item-to-item saccades reveal the collecting of this *relational* information, a point of view also endorsed by Salvucci and Anderson (2001), Thibaut et al. (2011), Hayes, Petrov, and Sederberg (2011), and others.

Thus, for both adults and children we considered their respective item-to-item saccade profiles (i.e., AB, AC, CT, etc.). We determined how well the various sets of these profiles allowed children to be distinguished from adults. We then compared LDA and SVM with three different kernels to determine how well each of these algorithms, when applied to various sets of item-to-item transitions, predicted whether the individual doing a problem was an adult or a child. We were particularly interested in making this prediction *as early as possible*, which is why we paid particular attention to item-to-item saccade profiles in the first third of the trial.

Predictions based on item-to-item saccades We looked at all of the item-to-item saccades (transitions) that were potentially relevant to solving the three types of A:B::C:D analogy problems given to participants. This set of transitions was *AB*, *AC*, *BC*, *BT*, *CT*, *CD*, and *TD*. Over the course of the trial, we counted the numbers of these item-to-item saccades that made up each scanpath. This gave us a "transition profile" for each participant and each trial. For example, suppose that for a given trial a child had eight AB transitions, two AC transitions, one AC transition, no BT transitions, 12 CT transitions, eight CD transitions, and four TD transitions; the child's {AB, AC, CT} transition profile for that trial would then be {8, 2, 12}, the {AB, TD} transition profile would be {8, 4}, and so on.

As we described earlier, there were three trial types: "Scene," "ABCD," and "ABCD-Scene." For each of these three trial types, we considered all possible sets of transitions (e.g., {CT}, {AB, BC}, {AB, CT, CD, TD}, etc., for a total of 127 different sets of transitions). We trained and tested an LDA classifier (Fisher, 1936) on each set of transitions using the LOOCV technique. In our case, this meant that for a given set of transitions (e.g., {AB, BC, TD}), and for the set of 45 participants, one participant was left out of the training set, and the LDA was trained on the other 44 participants. Then LDA attempted to predict whether the "left-out" participant was an adult or a child. We did this for all 45 participants and reported the percentage of correct predictions. This procedure was repeated for all 127 possible subsets of the set of seven item-to-item transitions (i.e., AB, AC, BC, BT, CT, CD, and TD). In this way, we were able to determine (1) which set of item-to-item transitions best predicted whether the participant was an adult or a child and (2) how good this prediction was.

We then ran an identical LOOCV procedure using a standard, two-class SVM classifier (Vapnik, 1995, 1998), using quadratic, polynomial (order 3), and radial-basis function (RBF) kernels. It is generally accepted that SVMs are some of the most powerful classifiers that exist. We also tested a standard backpropagation network with ten hidden units, learning rate = 0.005, momentum = 0.9, one output node, and a number of input nodes corresponding to the number of item-to-item transitions being tested. However, although we

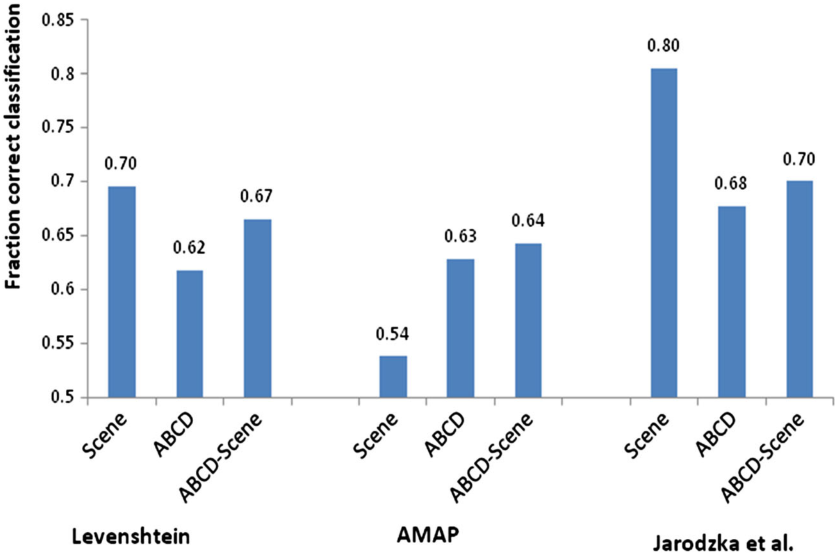


Fig. 5 A feedforward–backpropagation network trained on the points in the MDS maps derived from the scanpath-difference matrices for each of the three scanpath-comparison algorithms (Levenshtein, AMAP, and Jarodzka) and the three experimental conditions (Scene, ABCD, and ABCD-Scene). The AMAP algorithm is the poorest performer, and the Jarodzka et al. (2010) algorithm is clearly the best

found that its classification performance was acceptable, these networks were extremely slow, on the order of two orders of magnitude slower than the LDA and SVM algorithms. Therefore, we have not included them in this comparative analysis.

We considered only the transitions during the first third of each trial. The predictive powers of the six best transition profiles for each problem type are shown for LDA (Table 1a) and for SVM with an RBF kernel (Table 1b). We only show the results for the LDA classifier, which had the poorest classification performance, and the SVM classifier with an RBF kernel, which had the best. The runtimes of all classifiers were approximately the same.

Somewhat counterintuitively, prediction based on item-to-item saccade profiles is *better* if we look only at the first third of the trial than if we consider the whole trial. This is because, over the course of the entire trial, some item-to-item saccades for adults and children tended to balance out. For example, children might look at the CT transition more than adults in the first third of the trial, but less than adults in the final third. As a result, the overall numbers of CT transitions over the course of the whole trial could even out between children and adults and, for this reason, do not provide a good means of discriminating adults from children. On the other hand, the numbers of CT transitions in the first third of a trial are significantly different for children and adults and allow the two groups to be discriminated.

Finally, we looked at the overall numbers of item-to-item saccades for all participants during the first third of each trial for both adults and children for each of the three types of analogy problems. Children, in general, took longer than adults to do a given problem and, as a result, had a higher total number of saccades for each problem. For this reason, for each participant we normalized the data for each saccade type (i.e., AB, CT, etc.) by dividing his or her number of saccades for that saccade type by his or her total number of saccades (Table 2). We compared these normalized frequency values for each saccade type to the sets of transitions used by LDA and SVM to produce the best predictions as to whether an adult or child was doing an analogy problem.

Discussion

This article is not about analogy making per se. Rather, it concerns the quality of the classification methods and machine-learning techniques used to analyze eyetracking data produced in a study of the dynamics of analogy making. That said, it should be noted that these techniques, when applied to the eyetracking data generated by children and adults during analogy problem solving, allowed us to answer an outstanding problem in the field of analogy—namely, whether children use different strategies than adults when solving analogy problems.

Most importantly, in terms of methodology, we compared a number of widely used scanpath algorithms and found that the Jarodzka et al. (2010) algorithm is the most efficient one for examining scanpaths during analogy making. We also applied classic (LDA) and advanced (SVM) classification techniques to sets of the transitions making up scanpaths and demonstrated that these machine-learning techniques can be used to predict, well above chance, and in the first several seconds of a trial, whether the participant doing the problem is a child or an

Table 1 Correct-prediction probabilities using LDA (a) and SVM with an RBF kernel (b) for the six best sets of transition profiles for the three types of analogy problems

Scene		ABCD		ABCD-Scene	
P(Correct- prediction)	Transition profile	*P*(Correct- prediction)	Transition profile	*P*(Correct- prediction)	Transition profile
a. LDA					
.70	AB AC BT	.79	AC CT	.64	AB
.70	AB BT CT TD	.76	AB CT	.64	AB BC
.70	AB BT CT CD TD	.74	AC BT CT	.63	AC
.69	AC BC CT	.72	BC CT	.63	AC BT
.68	BT CD	.71	AB BC CT	.62	AC CD TD
.675	AC BC CD TD	.71	AB BT CT	.62	AC BC BT
b. SVM with RBF kernel					
.74	BT CD	.8	AB BC CT CD	.82	AB AC BC
.675	AC BC CD TD	.79	BC CT	.77	AB AC BC CT
.67	AB BT	.79	AC CT	.75	AB BC CT
.66	AC BC BT CD TD	.78	AB CT	.75	AB BT CT TD
.61	AC BC TD	.78	AB CT CD	.75	AB AC BC BT
.61	AC BC CD	.78	AB AC BT CT	.74	AC CT

adult. We also found that SVM with an RBF kernel produced the best adult/child predictions of the four classifiers tested. Finally, we found that certain subsets of item-to-item saccades predicted whether a child or an adult was doing a problem better than the full set of item-to-item transitions.

Table 2 shows the normalized differences (diff/max) between adults and children in the numbers of each type of transition for the three kinds of analogy problems in the first third of each trial. (The larger the value, the larger the difference will be between adults and children for a particular transition type.) Both LDA and SVM made use of the distinguishing differences between adults' and children's transition profiles during the first third of a trial to make their predictions. Thus, at least one of the transitions in a set of transitions used for prediction would, almost certainly, be a transition for which there was a large normalized difference between adults and children. Consider the subsets of transitions that resulted in the best predictions by the SVM–RBF algorithm for the three analogy problem types. For the scene analogies, SVM used the BT and CD transitions to produced the best prediction of whether a child or an adult was doing a problem (74 % accuracy). When we look at Table 2, we see that the two transitions that have the greatest normalized differences between adults and children are BT (.45) and CD (.61). For the ABCD analogy problems, the diff/max value of the CT transition (.93) is nearly twice as large as any other transition, and this transition is present in all six of the transition sets that produced excellent adult/child predictions (78–80 % accuracy). Finally, for transitions in the ABCD-scene problems, there is little variation between the normalized differences in Table 2 between adults and children. The top three transitions, based on their normalized differences, are BT (.38), BC (.33), and AC (.29). The six best distinguishing subsets, ranging in prediction accuracy from 74 to 82 % correct, all include at least one, and generally two, of these three transitions.

Table 2 Differences between the (normalized) numbers of transitions for adults and children, as compared to the maximum number of transitions

Diff/Max	AB	AC	AT	BC	BT	BD	CT	CD	TD
Scene	.13	.35	.21	.24	.45	.27	.08	.61	.08
ABCD	.29	.36	.46	.14	.55	.34	.93	.19	.37
ABCD-Scene	.10	.29	.24	.33	.38	.26	.06	.26	.28

These values were calculated as follows. Consider the BC transition for the ABCD problem type. For children, the normalized number (i.e., the fraction of the total number of transitions) of BC transitions was .28, and for adults this value was .24. The diff/max value in the table was obtained by taking the absolute value of the difference between these two values (i.e., .04) and dividing it by the maximum of both values (i.e., .28). Thus, we have (.28 – .24)/.28 = .14.

The point, in terms of methodology, is that the classification algorithms studied here provide an extremely powerful means of predicting whether a child or an adult is doing an analogy problem (or what the outcome of the trial will be), by spotting differences in strategies early in a trial. Analyses using LDA or SVM not only allow us to observe early-on differences in strategies that distinguish adults from children, but also reveal that the differences in their strategies also depend on the type of analogy problem being done. So, for example, with the ABCD analogy problems, both LDA and SVM show the CT transition to be important for adult–child classification, a fact that is borne out by the transition frequency counts in Table 2. On the other hand, these same analyses show that the CT transition is less important for the scene and ABCD-scene problems in predicting the age group (child/adult) of the participant.

Finally, it was not lost on us that these techniques could be applied to determining from the first third of a trial whether or not a correct answer would be given by a child for a particular problem. (Adults, for all intents and purposes, always answer the problems correctly, so we only ran this analysis with children.) Although we do not present the data in this article, we ran a second experiment, very similar to the one described above, in which we looked at this. These results are reported in French and Thibaut (2014). We found that by looking at a set of two item-to-item transitions, {AB, CT}, in the first 3 s of a trial, we could predict with an accuracy well above chance (62.5 %) whether or not a child would answer a given problem correctly.

The bottom line is that scanpath-comparison algorithms and the machine-learning techniques that accompany them are powerful tools to study the dynamics of analogy making. In building models of analogy making, we want to know what the models predict and how they make those predictions. Although the tools presented in this article are more involved with prediction than with explanation, the two are hardly unrelated, especially when we know the bases of the predictions. Our overarching goal has been to point reseachers in analogy making toward tools and analysis techniques that will allow them to better study the dynamics of how people solve analogy problems.

Conclusion

Eyetracking technology has come of age. Equipment that, as little as a decade ago, cost tens of thousands of dollars can now be purchased for several hundred. More and more researchers in the behavioral sciences are using this technology to probe the mechanisms underlying diverse cognitive skills, in general, and analogy making, in particular. By comparing a number of scanpath-comparison algorithms and machine-learning techniques that can be applied to raw data generated by eyetrackers, we hope to have pointed researchers to tools that will best serve them as they attempt to study the dynamics of analogy making.

Author note This research has been supported by joint ANR–ESRC Grant No. ORA 10-056 GETPIMA to the first author, a Fondation Fyssen grant to the second author, and French ANR Grant No. 10-BLAN-1908-01 to the third author.

References

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4,* 40–79.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8,* 205–238.

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling*. New York: Chapman & Hall.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827–1837. doi:10.1016/0042-6989(95)00294-4

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1,* 269–271.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188.

French, R. M., & Thibaut, J.-P. (2014). Using eye-tracking to predict children's success or failure on analogy tasks. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 2222–2227). Austin: Cognitive Science Society.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70,* 320–328.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170. doi:10.1207/s15516709cog0702_3

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 225–277). New York: Cambridge University Press.

Gentner, D., & Smith, L. (2012). Analogical reasoning. In *Encyclopedia of human behavior* (2nd ed., Vol. 1, pp. 130–136). Amsterdam, The Netherlands: Elsevier Inc.

Glady, Y., Thibaut, J. P., & French, R. M. (2013). Visual strategies in analogical reasoning development: A new method for classifying scanpaths. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmith (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2398–2403). Austin: Cognitive Science Society.

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition, 15,* 20–35. doi:10.1080/13506280600871891

Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale: Erlbaum.

Goswami, U., & Brown, A. L. (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition, 36,* 207–226. doi:10.1016/0010-0277(90)90057-Q

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale: Erlbaum.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision, 11*(10), 10:1–11. doi:10.1167/11.10.10

He, P., & Kowler, E. (1992). The role of saccades in the perception of texture patterns. *Vision Research, 32,* 2151–2163.

Hofstadter, D. R. (2001). Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–535). Cambridge: MIT Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York: Oxford University Press.

Holyoak, K. J., Gentner, D., & Kokinov, B. (2001). The place of analogy in cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 1–19). Cambridge: MIT Press.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In C. H. Morimoto, H. Istance, A. Hyrskykari, & Q. Ji (Eds.), *ETRA '10: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 211–218). New York: ACM. doi:10.1145/1743666.1743718

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8,* 441–480. doi:10.1016/0010-0285(76)90015-3

Lachenbruch, P. A. (1967). An almost unbiased method for the probability of misclassification in discriminant analysis. *Biometrics, 23,* 639–645.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 45,* 251–266. doi:10.3758/s13428-012-0226-9

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

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11,* 157–178. doi:10.1163/156856897X00177

Miller, R. G. (1974). The jackknife: A review. *Biometrika, 61,* 1–15. doi:10.1093/biomet/61.1.1

Nodine, C. E., Carmody, D. P., & Kundel, H. L. (1978). Searching for Nina. In J. Senders, D. F. Fisher, & R. Monty (Eds.), *Eye movements and the higher psychological functions* (pp. 241–258). Hillsdale: Erlbaum.

Ouerhani, N., von Wartburg, R., Hugli, H., & Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis, 3,* 13–24.

Rajashekar, U., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2008). GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing, 17,* 564–573. doi:10.1109/TIP.2008.917218

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94,* 249–273. doi:10.1016/j.jecp.2006.02.002

Rumelhart, D., McClelland, J., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge: MIT Press.

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path-mapping theory. *Cognitive Science, 25,* 67–110.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B, 36,* 111–147.

Thibaut, J.-P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development, 38,* 10–26.

Thibaut, J.-P., French, R. M., Missault, A., Gérard, Y., & Glady, Y. (2011). In the eyes of the beholder: What eye-tracking reveals about analogy-making strategies in children and adults. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Expanding the space of cognitive science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 453–458). Austin: Cognitive Science Society.

Thibaut, J.-P., French, R., & Vezneva, M. (2010a). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review, 17,* 569–574. doi:10.3758/PBR.17.4.569

Thibaut, J.-P., French, R., & Vezneva, M. (2010b). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology, 106,* 1–19. doi:10.1016/j.jecp.2010.01.001

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17,* 401–419.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Berlin: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley.

Wagner, R., & Fischer, M. (1974). The string-to-string correction problem. *Journal of the ACM, 21,* 168–178.

Woods, A. J., Göksun, T., Chatterjee, A., Zelonis, S., Mehta, A., & Smith, S. E. (2013). The development of organized visual search. *Acta Psychologica, 143,* 191–199. doi:10.1016/j.actpsy.2013.03.008