Using eye-tracking to predict children's success or failure on analogy tasks

Robert M. French and Jean-Pierre Thibaut

{robert.french, jean-pierre.thibaut}@u-bourgogne.fr
LEAD-CNRS, UMR 5022, University of Burgundy, Pôle AAFE – Esplanade Erasme
21065 DIJON. FRANCE

Abstract

We use eye-tracking data, analyzed by a neural network and by Linear Discriminant Analysis (LDA), to study the *temporal dynamics* of children's analogy making. We determine how well the number of item-to-item saccades while solving an analogy problem predicts whether or not a child will correctly answer the problem. For the A:B::C:D visual analogy problems, by the first third of the trial we can tell with 64% accuracy whether or not the problem will be answered correctly. Two-thirds of way through the trial, we can predict with 82% accuracy the answer that will be given. By looking only at the final third of the trial, we can predict with up to 90% accuracy what the child will do. Average gaze times at the Target and Distractor items have the same predictive power as the item-to-item saccade information.

Keywords: Analogical reasoning; development; eye-tracking; strategies, prediction in analogy making

Introduction

The centrality of analogical reasoning to human cognition is not open to debate (Gentner & Smith, 2012; Holyoak, 2012). On the other hand, there are many open questions surrounding how analogy making occurs and how the ability to do analogies develops over time.

Very little work has been done on the *dynamics* of solving analogy problems. The vast majority of experiments involve selecting particular items as the solution to an analogy problem. This is a static approach to analogy making and, by definition, cannot address the issue of how a strategy evolves during an attempt to solve a problem. For a number of years, we have used eye-tracking technology to study analogy-making strategies, particularly in children. As an analogy problem is being solved, the information sought and manner in which it is sought can be elucidated by the visual strategies by the problem solver. The amount of attention paid to a particular item and the gaze-fixation on that item have been shown to be highly correlated, in particular, for complex stimuli (Deubel & Schneider, 1996; He & Kowler, 1992). In addition, the fixation time associated with a given item correlates with its informativeness (Nodine, Carmody, & Kundel, 1978). All of this argues in favor of using eye-tracking technology to study analogy-making strategies.

There are only a small number of eye-tracking studies involving analogy-making in adults (e.g., Gordon & Moser, 2007) and even fewer in the area of the development of analogy-making abilities in children. One of the first developmental analogy-making studies was one by

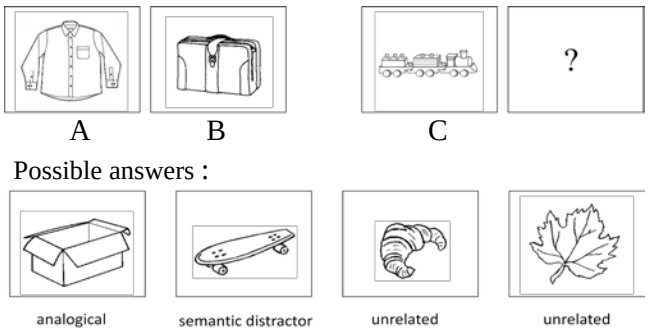

Figure 1. The type of A:B::C:? analogy problems used in the present study

Thibaut, French, Missault, Gérard & Glady (2011) in which they established that children's and adults' analogy-making strategies differed. They showed, in the context of the A:B::C:? paradigm, that "adults looked more at A and B than at C and Target and that they start with A and B before looking at C and D, [whereas children] spent significantly more time than adults on C and the Target item (or distractors) and less on A and B."

More recently Glady, French & Thibaut (2013) focused not on looking times, but rather on full gaze-saccade paths (called *scanpaths*) of adults and children on three different sets of analogy problems. They developed a novel technique for analyzing these scanpaths that involved: i) modifying an algorithm developed by Jarodska, Holmqvist, & Nyström (2010) that calculated a "distance" between any two scanpaths, ii) using a classic multidimensional scaling algorithm to represent each scanpath as a point in R^2 in such a way as to optimally preserve the distances calculated by the Jarodska et al. algorithm, and iii) using a multi-layered perceptron with a Leave-One-Out cross-validation procedure to classify and test these points according to whether they corresponded to scanpaths generated by adults or by children. By means of this technique they were able to demonstrate that there was, indeed, a clear difference in the dynamics of strategies used by children and adults in solving analogy problems of the form A:B::C:D. However, this technique revealed only that adults' and children's analogy-making strategies were different. The present paper is an attempt to answer a closely related question using eye-

tracking methodology that expands on this previous work. The question is the following:

> *To what extent can we determine whether a child will answer a given analogy problem correctly or incorrectly based solely on an analysis of his/her patterns of eye movement at various times during the trial?*

In the present paper, we divide each scanpath into three identical time segments, which we label "initial", "middle", and "late" and which correspond to the 1st, 2nd, and 3rd thirds of the full time interval over which the scan occurred. Within each of these three segments we will determine the predictive value of the child's pattern of item-to-item transitions (e.g., how often does he/she saccade from item A to item B, from C to the Target item, etc?).

In what follows we will first present the experiment that we ran and graphically present the item-to-item transition data from that experiment. We will then analyze by means of a multi-layer perceptron (and a linear discriminant analysis) the predictive value of this data in determining, as early as the initial time segment, the answer that children will give on the analogy problem they are attempting to solve.

Experiment

Methods

Participants

Thirty-nine 5-year-old children (M = 5;7, age range, 4;9-6;2), thirty-seven 8-year-old children (M = 8;8, range 8;0 – 8;10), 23 adolescents (M = 13;9, range 13;3-14;3) and 20 adults (M = 21;7, range 19;4-25;4) participated in the experiment. We tested a larger number of children because we expected a greater loss of eye-tracking data for the younger groups. In each of the time slices, if there were no recorded item-to-item transitions (i.e., the child was not looking at the screen), we removed the child from the data for that slice. The data removed was approximately the same for the younger and older children. We had to remove 25%, 28%, and 30% of the data for the initial, middle and final time slices, respectively. Only children were considered for this experiment, since for essentially all adolescents and adults responded correctly to all analogies.

Informed consent was obtained from the children's parents.

Materials

The experiment consisted of a total of 14 trials divided into 2 practice trials and 12 experimental trials. Each of the two distractor conditions (No or 1 semantic distractor) consisted of 6 trials. Each trial contained 7 drawings -- namely, the items corresponding to the A, B and C items and the 4 drawings that were shown as the solution set, including the analogical match (hereafter, "Target" or, simply, "T"). In the one-semantic-distractor case, there was one semantically related distractor ("SemDis", or simply "D"), and two distractors that were semantically unrelated to C ("UnDis", or simply "N"). (Note: There are two UnRelated Distractors in each problem. We averaged the looking times to these two distractors and used this value for N.) In the No-Semantic-Distractor condition the three distractors were semantically unrelated to C. In the scanpath analysis described in this paper, however, we consider only the trials in which there was a semantic distractor. The reason for this was that when there was no semantic distractor, all children solved the analogy problem correctly. We are interested in predicting whether or not a child will answer the problem correctly based on his/her gaze pattern and therefore a semantic distractor was necessary to induce the children to answer some problems incorrectly.

The experiment was run with the E-prime® software. We used a Tobii T120 to record gazes.

Procedure

Two experimenters saw the children individually at their school in a quiet room or, for the adults, in our laboratory. Participants were seated in front of the Tobii screen with their eyes at a distance of approximately 40 centimeters. For each participant, the experiment started with a calibration phase which followed the protocol specified for the apparatus.

The participants were then shown picture cards of each of the items used in the experiment and were asked to give their names. When they did not know an item's name, they were asked to describe it functionally. The children knew 96% of the names and when they did not, in most cases, they were able to give a description showing that they knew the stimulus. The percentage of cases where children could not name the item or could not provide a correct description (functional or contextual) was 1%. In these cases, the experimenter gave the children the missing information.

Each trial began when the experimenter pressed the space-bar. The 7 stimuli for each trial were displayed simultaneously on the screen. The A:B pair and the C item were shown in an array with the first two items grouped together to the left of the screen. The C item was alone on the right of the screen and next to C there was a box with a question mark. The four solution items were displayed on a separate row, beneath the A B C ? row (see Figure 1). Participants were asked to point to the item in the lower row that best completed the series of items in the upper row (cf. Goswami & Brown, 1990). The first two trials were training trials and participants received feedback. In these two training trials the experimenter explained why the Target was the correct solution and incorporated the relation

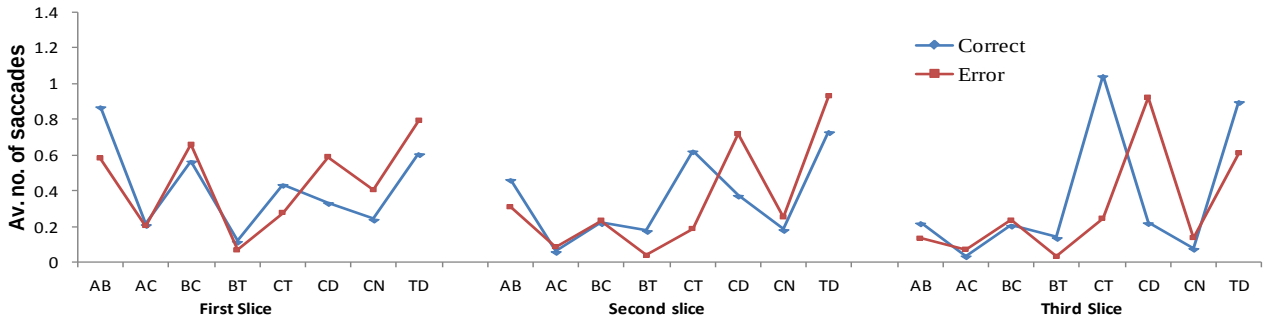

Figure 2a: Transition counts for Correct answers vs. Errors as a function of Time-Slice and Transition-type.

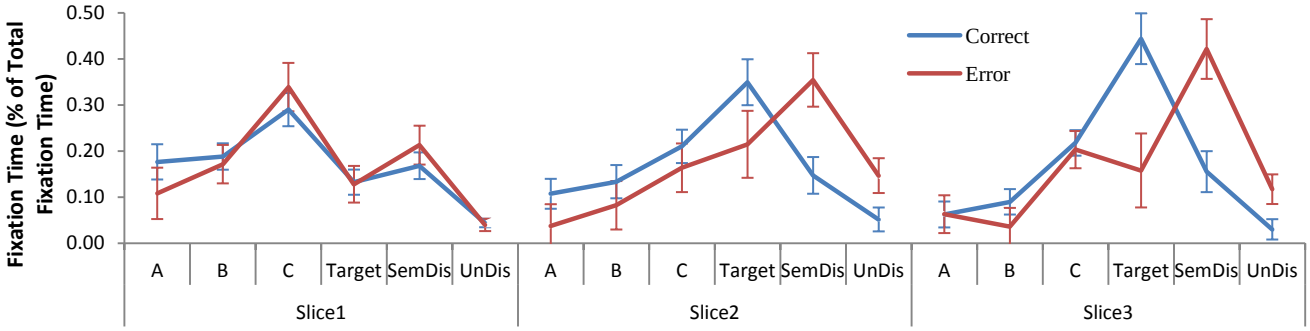

Figure 2b. Item gaze times for Correct answers vs. Errors as a function of time-slice and item type.

holding in the A-B pair in his/her demonstration. In the experimental trials, the experimenter gave no further information to children. For each experimental trial, reaction times were recorded by the experimenter. Participants were instructed to point to the stimulus on the screen corresponding to their choice "as soon as they had found the solution". They were told that they were to point to only one stimulus per trial. The experimenter stopped timing the participant when he/she pointed to a solution.

At the conclusion of the experiment, for all the experimental trials, children's understanding of the semantic relation between A and B and between C and D was assessed. They were shown the A:B pairs and were asked *why* the two items of each pair went together. The same was true for the C-D pairs.

Results

For the sake of clarity, we will consider only a small number of statistical results from this experiment. The dependent measure is the average number of transitions between items (i.e., saccades from one item to the other) for the following item pairs: A-B, A-C, B-C, B-Target, C-Target, C-SemDis, C-UnDis, and Target-SemDis. These are shown in Figure 2a. One effect is of particular importance and that is the three-way interaction between Response-Type x Time-Slice x Transition-Type. This interaction is highly significant: $F(14,994) = 3.38$, $p < 0.0001$; $\eta^2 = .045$. In other words, some of the individual item-to-item transitions really do count in distinguishing between correct and incorrect responses. Not surprisingly, over the course of the trial, for children who answered the problem correctly the frequency of C-Target transitions increased, whereas for those children who answered incorrectly, the frequency of C-Distractor transitions increased. We see the same evolution when we look at item gaze times (Figure 2b). A relatively detailed item-to-item transition analysis allows us to predict with considerable accuracy whether or not a child will answer a given problem correctly.

Predicting Children's Answers

Thibaut et al. (2011) showed that children tend to have different item gaze-time profiles than adults for the same analogy problem, leading to the conclusion that adults and children process analogies differently. Glady et al. (2013) reinforced this result using the full scanpaths of adults and children. We wondered whether scanpath analyses could also be used to whether a child would answer a problem correctly or incorrectly. In particular, we wanted to know the predictive quality of the pattern of item-to-item saccades at various points during the trial. (Note: these analyses were not done for the adult data because adults, unlike children, answered all of the analogy problems correctly.)

How the predictions were made

We used two widely used and relatively simple methods to calculate how well the observed item-to-item transition patterns in each time slice predicted the final response outcome. These were i) a feedforward-backpropagation

(FFBP) neural network (Rumelhart & McClelland, 1986) and ii) Linear Discriminant Analysis (LDA, Fisher, 1936; Rao, 1948). The goal of this paper is not to demonstrate the maximum predictive power of item-to-item transition patterns. For this there are more powerful classification algorithms (e.g., support vector machines, Vapnik, 1995). Rather, we wished to demonstrate that very early in solving an analogy-making problem, children's correct (or incorrect) response can be predicted at well above chance levels by observing their item-to-item saccading patterns. FFBP neural-network classifiers and standard LDA techniques are sufficient to clearly show this. We used two very different classification methods in order to ensure that the predictions were not an artifact of the classification technique being used.

Data classifiers—whether LDAs, FFBP neural networks or some other classification system—are designed to associate items with the categories to which they belong. So, for example, if a classifier learns to associate certain feature values with the category "dog" and other feature values with "cat", when it encounters a new cat (or a new dog) it should be able to correctly classify it. In other words, the classifier, based on the cats and dogs that it has been trained on, can correctly generalize to new exemplars of each category.

Data classification can also be used to indicate how well a given set of feature values will correctly predict the category membership of the item associated with those features. There is a standard train-and-test technique for doing this. The classifier is first trained on a large, randomly selected subset of the original data—for example, 75% or 80% of items in the dataset—and is then tested on the remaining data. This procedure is repeated many times, each time randomly dividing the data into a training set consisting of 75-80% of the data, and a test set comprised of the remaining data. Then average the classification success rates over the number of times the procedure was repeated.

Predictions from the data

We consided the data from the experiment described above and the results of which are shown in Figure 2. To analyse the data we first used a neural network and then performed a linear discriminant analysis on the same data.

The data input to both classifers consisted of the average number of item-to-item saccades for the different pairs of items shown in Figure 2. The neural network used had a variable number of inputs, depending on the transition pairs used. So, for example, if we considered only the transition pairs **CT** (i.e., between C and Target), **CD** (i.e., between C and the semantic distractor), and **TD** (i.e., between the Target and the Semantic Distractor), the neural network would have 3 inputs, one for the number of saccades associated with each of those three transitions. There were always 10 hidden units and one output unit. The learning rate was 0.001, with a momentum of 0.9 and a Fahlman offset of 0.1. A standard tanh squashing function, with a temperature of 1 was used. The network was allowed to run for a maximum of 1000 epochs. For each set of transitions, we ran the network 30 times, each time training the network on 75% of the data and testing it on the remaining 25%. For each run, we recorded how accurately the network was able to predict the real outcomes (correct/incorrect answer) on the 25% of problems that it had not been trained on.

We began by considering all eight of the transition pairs (ie., AB, AC, BC, BT, CT, CD, CN, TD). The question was how well the full pattern of transistions predicted the output (i.e., correct/incorrect response) based on the average number of saccades for each of these transitions during the Initial, Middle and Final time segments of the trial?

We found that using all of the transitions, by the end of the first time slice (i.e., after approximately 4.25 seconds of a 13-second trial, 13 seconds being the mean RT for children), we could predict the correct/incorrect outcome of the trial with a 64% accuracy. The transition information in the Middle time slice was even more informative, predicting the outcome of the trial with a 68% accuracy. In the final time slice the item-to-item saccade information allowed us to predict the outcome of the trial with an accuracy of 86%. (See Figure 3.)

To ensure that these results were not an artifact of the classification method or of the item-to-item transition data, we performed the same analysis, using the pattern of looking times at the various items comprising the problem. Thus, input to the network consisted, not of the average number of transitions between items, but, rather, the looking times at A, B, C, Target (T), SemDis (D), and UnDis (N).

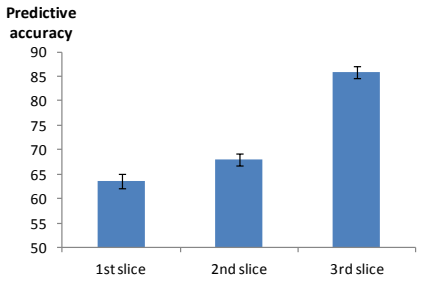

Figure 3. The predictive power of the full set of item-to-item transitions: AB, AC, BC, BT, CT, CD, CN, TD. (SEM error bars)

Using this information, we performed the same analysis as with the FFBP neural network whose parameters are indicated above. As can be seen in Figure 4, both of these measures, the average number of item-to-item transitions and looking times at the various items making up the analogy problem, give very similar results. In other words, the results shown in Figure 3 do not an artifact of having used the number of item-to-item transitions as our dependent measure. The accuracy of these predictions is also borne out by using as our dependent measure looking times at the individual items making up the problem. The principles of LDA classification are significantly different from those underlying FFBP network classification. For this reason, we also applied an LDA classifier on the item-to-item transition data to ensure that the results in Figure 3

were not an artifact of the classification method. We used the same set of transitions and ran the LDA classifier 30

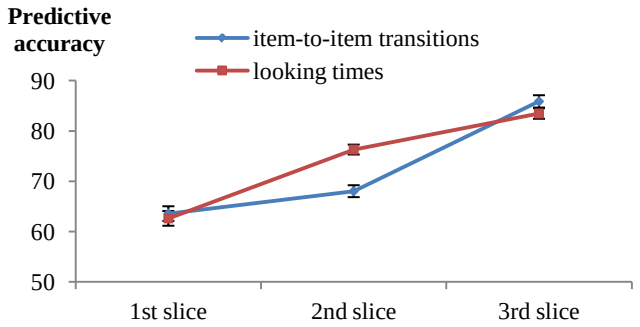

Figure 4. Accuracy of Correct/Incorrect response predictions based on item-to-item transitions and on looking times at the items of the problem.

times. As with the FFBP classifier, we used a 75%-25% train-and-test split of the data. The results using the LDA classifier were essentially the same as those obtained by the FFBP neural network classifier (Figure 5).

It is worth noting that, even though an LDA is not as powerful as a FFBP neural network in what it can and cannot classify, it is *much* faster. For example, in the present problem LDA was faster than the FFBP neural network by two orders of magnitude. And, as can be seen from Figure 5, there is no classification performance difference.

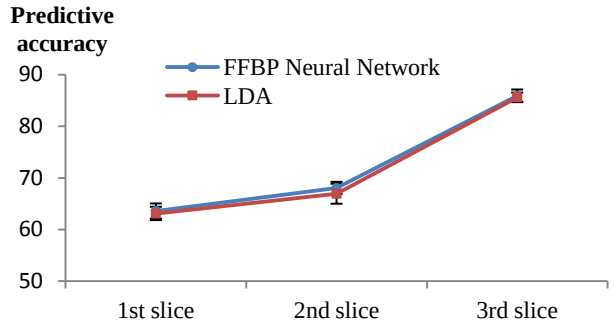

Figure 5. Comparing the FFBP neural network classifier and an LDA classifier.

Predictive accuracy of other sets of transitions

We also explored the predictiveness of different subsets of the item-to-item transitions. For this, we used an LDA classifier. As can be seen in Figure 6, sets multiple transitions involving C are the most predictive of the outcome of the trial. In fact, by looking at only two item-to-item transitions, those between C and the Target item and those between C and the Semantic Distractor, we can achieve a very high degree of prediction accuracy. On the other hand, transitions not involving C or any of the potential target items are not particularly informative in predicting the final outcome of the trial. It is worth noting that prediction value of CT transitions increases over time, while the opposite is true for TD transitions.

	1st slice	2nd slice	3rd slice
CT, CD, TD	0.62	0.68	0.9
CT, CD	0.625	0.72	0.9
CD, TD	0.61	0.6	0.84
CT, TD	0.62	0.67	0.72
CD	0.66	0.6	0.65
CT	0.5	0.54	0.6
TD	0.6	0.56	0.51
AB, AC, BC	0.55	0.58	0.58
AB	0.47	0.47	0.56

Figure 6. Predictive accuracy of various subsets of item-to-items transitions (LDA classifier, average of 100 train-and-test repetitions)

Combining the First and Second Slices

Rather than looking at each of the individual time slices (i.e., "Initial", "Middle", "Final"), in our final analysis we looked instead at the first two-thirds of the trial. In other words, we combined the first two slices. The children's average item-to-item transition profile for the first two-thirds of the trial is shown in Figure 6. Their average looking-times at the different items over the first two-thirds of the trial is shown in Figure 7.

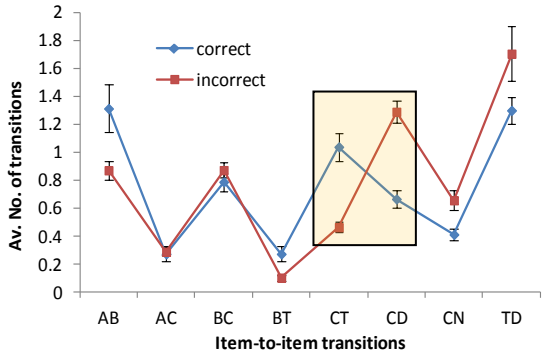

Figure 6. Average number of item-to-item transitions after two-thirds of the trial. The key transitions **CT** and **CD** are highlighted. (SEM error bars)

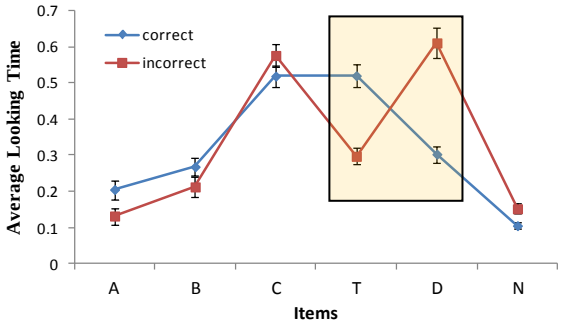

Figure 7. Average item looking times after two-thirds of the

trial. The key items, the Target (T) and the Semantic Distractor (D), are highlighted. (SEM error bars)

Transitions or items	After 2/3 of the trial
CT, CD	0.76
T, D	0.82

Table 2. Prediction accuracy (LDA analysis over 100 train-and-test runs) for the key item-to-item transitions, **CT** and **CD**, and the key items, **T** and **D**.

Our prediction about how a child would answer a given problem can, therefore, be based on combining the information in Table 2 with the graphs in Figures 6 or 7. For example, if after two-thirds of the trial a child has transitioned significantly more from C to the Target (T) than from C to the Semantic Distractor (D) (Figure 6), we can say with a probability of 0.76 (Table 2) that he/she will answer the problem correctly. If, on the other hand, after two-thirds of the trial, the child has transitioned significantly more between C and the Semantic Distractor than between C and the Target item, then we can say with a probability of 0.76 that he/she will answer the problem incorrectly.

The same holds for item looking times. If, after two thirds of the trial, the child has looked at the Target item (T) significantly longer than the Semantic Distractor (D) (Figure 7), then we can say, with a probability of 0.82 (Table 2) that he/she will answer the problem correctly. And, as before, if the child has looked significantly longer at D than at T over the course of the first two-thirds of the trial, the probability of answering the problem incorrectly is 0.82.

Discussion and Conclusion

This goal of this paper has been to explore the temporal dynamics of analogy making in children. Specifically, we examined the question of just how much the patterns of item-to-item gaze transitions could tell us about whether or not the child would give the correct answer to an analogy problem.

It turns out that very early, in the first third of a given trial, the predictive accuracy of particular patterns of transitions is well above chance. By two-thirds of the way through the trial (Table 2), prediction accuracy rises to around 80%. And by the last third of the trial, we can predict with almost 90% accuracy how the child will respond based on these item-to-item gaze transitions (and also, as we have shown, on their looking times at the items themselves).

The techniques that we have used to classify children's performance on A:B::C:D analogy problems based on their profiles of looking times or their number of item-to-item saccades give us a better understanding of the temporal dynamics of the children's decision process. While it is true that the analyses in this paper apply to a specific kind of analogy problem — the A:B::C:? paradigm — there is no reason to believe that this kind of analysis could not be used in broader, more ecological analogy-making contexts. More than anything, these preliminary results paper argue strongly for the use of eye-tracking methodology to study the temporal dynamics of analogy making, allowing us to better understand how the solving of (or the failure to solve) an analogy problem unfolds over time.

In this paper we have seen just how rich and informative this kind of analysis of analogy making can be.

Acknowledgements

This research has been supported by French ANR Grant 10-BLAN-1908-01-Anafonex to the second author and a joint ANR-ESRC grant 10-056 GETPIMA to the last author.

References

Deubel, H., & Schneider, W. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. Vision research, 36, 1827–1837.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics 7 (2): 179–188.

Gentner, D., & Smith, L. (2012). Analogical Reasoning. *Encyclopedia of Human Behavior* (2nd ed., Vol. 1). Elsevier Inc.

Glady, Y., French, R. M., and Thibaut, J. P. (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 Thirty-fifth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2398-2403.

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition*, *15*(1), 20–35.

Goswami, U. & Brown, A.L. (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. Cognition, 36, 207-226.

He, P., & Kowler, E. (1992). The role of saccades in the perception of texture patterns. *Vision research*, *32*(11), 2151–2163.

Holyoak, K. J. (2012). Analogy and Relational Reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*. New York, NY: Oxford University Press.

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*. Hillsdale, NJ: Erlbaum.

Rao, R. C. (1948). "The utilization of multiple measurements in problems of biological classification". Journal of the Royal Statistical Society, Series B 10 (2): 159–203.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: MIT Press

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. *Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society* (pp. 453–458).

Vapnik, Vladimir N.; *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.