

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

A method for quantifying focused versus overview behavior in AOI sequences

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/75859> since 2017-03-29T15:00:03Z

Published version:

DOI:10.3758/s13428-011-0104-x

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



UNIVERSITÀ DEGLI STUDI DI TORINO

The final publication is available at Springer via <http://dx.doi.org/10.3758/s13428-011-0104-x>

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

A method for quantifying focused versus overview behavior in AOI sequences

Kenneth Holmqvist · Chiara Andrà ·
Paulina Lindström · Ferdinando Arzarello ·
Francesca Ferrara · Ornella Robutti · Christina Sabena

Abstract We present a new measure for evaluating focused versus overview eye movement behavior in a stimulus divided by areas of interest. The measure can be used for overall data, as well as data over time. Using data from an ongoing project with mathematical problem solving, we describe how to calculate the measure and how to carry out a statistical evaluation of the results.

Keywords Moving window · AOI sequence · Transitions matrix · Autocorrelation · Global vs. local scanning · Ambient vs. focal scanning

Introduction

This article describes a new measure: *transition sequences between areas of interest*, or in short, *TS-AOI*. This measure is used for analyzing data coming from the dynamic behavior of eye-movements when areas of interest (AOIs) are explored.¹ In this study, the AOI is a limited area on a stimulus encapsulating a word in a text or an object in a picture. The new measure combines two features that existing measures have not combined to date. First, it classifies subsequences of eye-movement data as *focused* behavior (looking within or between a few positions) versus *overview* behavior (looking at many or all positions in a sequence).

¹ The AOI is used as in established eye-tracking research literature.

K. Holmqvist (✉) · P. Lindström
Lund University, Lund, Sweden
e-mail: Kenneth.Holmqvist@humlab.lu.se

C. Andrà · F. Arzarello · F. Ferrara · O. Robutti · C. Sabena
Turin University, Turin, Italy

Second, it represents gaze position data with AOIs that can be given semantic meaning in relation to the experiment at hand.

The need for the new measure was revealed in a project investigating mathematical problem solving, where we were interested in dynamic changes in eye movement behavior *over time*, using an AOI representation of stimuli. We specifically wanted to analyze the transitions among the areas that contained relevant information for solving the given task—for example, to know how participants change between an overview and a focused behavior over time, since these two behaviors signify two cognitive processes that could be useful for determining what a participant is doing over time in a task. In mathematical problem solving, focused inspection of a few AOIs could be a comparison between two mathematical objects, while a sequence over many AOIs could be overview scanning or search for a specific element.

Focused versus overview eye-movement behavior is a well-known distinction in eye-tracking research. Early observations by Buswell (1935) found that art viewing involves two kinds of eye-movement behavior, either short fixations over the whole painting or longer fixations in a delimited area of the painting. They correspond to overview and focused eye-movement, respectively. More recently, these two behaviors were reinvestigated by Unema, Pannasch, Joos, and Velichkovsky (2005), using the terms *ambient* versus *focal* search. Groner, Walder, and Groner (1984) and Zangemeister, Sherman, and Stark (1995) also studied these eye movement characteristics, using the terms *global* versus *local* search.

Sequences of AOI hits are often analyzed using transition matrices (Ponsoda, Scott, & Findlay, 1995), Markov models (Simola, Salojärvi, & Kojo, 2008), or varieties of the string edit similarity measures (Choi, Mosley, & Stark, 1995). Change over time and space has also been studied using proportion-over-time analyses (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) and time series analysis (Uttal, 1983).

However, as we will argue next, although all these methods have useful qualities, they do not combine the use of AOIs with analysis of subsequences of focused and overview behaviors. We then show not only that the new unique AOI measure is clearly distinct from already existing measures, but also that it is a general measure sensitive to small differences in overall visual behavior during complex tasks. And it can be easily integrated with other measures, in order to have a deeper and broader look at data from eye-movement behavior. This measure lends itself well to a range of domains where the aim of the study is to examine eye-movement behavior over time. In usability research, human factor, and advertisement studies, AOI sequences are commonly used in data analysis. Our own application is problem solving and educational psychology, but psycholinguistic visual world and general real-world studies that use AOIs and investigate sequences are equally likely to benefit from implementing this measure.

Method

Participants Forty-six Swedish university students participated in the study. Students had different knowledge backgrounds in mathematics. Twenty-four had no previous university courses in mathematics, and 22 had one year of mathematics studies at an engineering faculty. We later divided the participants into two equally large groups on the basis of their overall score (number of correct answers)–low-ability and high-ability participants.

Materials All 43 mathematical stimuli had exactly five AOIs, which were the same distinct objects as in Fig. 1. There was always one text, graph, or formula, as in this case, which is referred to as the input (labeled “I”), and four alternatives (labeled “A”–“D”), which are all of the same type but different from the input: 15 presented a single formula to the right on the screen, with four text alternatives to the left on the screen; 12 presented a single graph occupying the top two thirds of the screen, with four alternatives containing text under the graph; 16 presented a text occupying the top one third of the screen, with four alternatives containing formulas below the text to the right on the screen. The content of the AOIs varied in size depending on the content; for example, a short formula resulted in a smaller AOI, whereas a graph required a larger AOI. There were no overlaps between the AOIs. All stimuli were sized to fit the screen specified in the Apparatus section.

Apparatus We used an SMI HiSpeed eye tracker with a sampling frequency of 1250 Hz to collect transitions, which are eye-movements between the different AOIs. In a sequence such as IAIB, a participant first looks at the

$$\text{I=INPUT}$$

$$y = \frac{7}{2-x} - 3$$

A
y skär den vertikala axeln i en punkt med negativt y-värde.

B
y skär den horisontella axeln i en punkt med negativt x-värde.

C
y har en vertikal asymptot $x = 3$.

D
y har en horisontell asymptot $y = 3$.

Fig. 1 Mathematical-problem-solving task: Find the correct written description of the function in the formula (I = input). Translations of the four alternatives are A, “y cuts the vertical axis in a point with negative y-value”; B, “y cuts the vertical axis in a point with negative x-value”; C, “y has a vertical asymptotex = 3”; D, “y has a vertical asymptotey = 3”

input, then at alternative A, then at the input, and finally at alternative B. Stimuli were presented on a computer screen (380 × 304 mm) with E-Prime, at a distance of 56 cm.

Procedure The participant was seated in front of the eye-tracker and placed his or her head in the chin rest of the tower-mounted system to stabilize viewer position. We calibrated the eye tracker and presented a practice block in order to familiarize the participant with the task, which was to determine which of the four alternatives (A–D) was an accurate representation of the single representation (I) and corresponded to the single representation according to some rule or feature. All stimuli were thereafter presented in random order. Each participant took his or her own time to solve the task, but each stimulus was presented for a maximum of 40 s, which was pre-tested as a sufficiently long time to solve the task. Participants chose their answer by clicking on it with the mouse.

The problem to be solved

Figure 1 shows 1 out of 43 mathematical-problem-solving tasks from the study that will exemplify our measure.

In this study, we wanted to answer questions such as the following: When participants work with these mathematical tasks, do they mostly go back and forth between a couple of AOIs (which we call *focused behavior*), or do they very often circle all five AOIs in a sequence (which we call *overview scanning*)? Do they sequence all five AOIs in the early phase and only then make transitions back and forth between a few AOIs?

We were interested in such an analysis because it could provide information about changes in problem-solving behavior over time. In our particular application, we want to use the results in future intervention studies to examine how students can be supported when reading mathematical representations.

The existing measures

Here, we present seven already existing measures, or groups of measures, and evaluate them against the requirements that the experimental design in our problem-solving study place on our proposed measure. That is, we examine how the measure accounts for the two properties of interest: representation of position, using AOIs, and the frequency of focused versus overview sequences. We have not included pure position comparisons and spatial dispersion measures that obviously do not take sequences into account—for instance, the Mannan similarity measure (Mannan, Ruddock, & Wooding, 1996) or the Kullback–Leibler distance measure (Tatler, Baddeley, & Gilchrist, 2005).

Local versus global The degree to which a scanpath is *local* is operationalized as the proportion of saccades with an amplitude below a certain threshold (Groner et al., 1984), which is typically selected to be around 1.5° . Local scans (short saccades) are argued to correspond to detailed inspection, while global scans (long saccades) reveal overview looking. There are three drawbacks of the local versus global measure for our problem-solving task. (1) It cannot take AOI data as input. (2) Detected local short amplitudes are likely to simply reflect close inspection inside an AOI (as in Fig. 2). (3) Moreover, the measure does not involve any analysis of sequences.

Ambient versus focal The distinction between *ambient* and *focal* processing combines saccadic amplitudes with fixation durations: An ambient scan is recognized by long saccades and short fixations, while a focal scan involves short saccades and long fixations. Unema et al. (2005) showed that when viewing pictures, participants start with an ambient overview scan and only then scan the picture elements focally. From the perspective of our study, drawbacks with this measure are, again, that it is not able to accept AOI data and that detected focal processing can be internal to AOIs. Sequences are not involved in the analysis.

Proportion of time analyses When the proportion of participants that look at a specific AOI is plotted over time, the resulting graph can be used to study how quickly and in what order AOIs are seen (Tanenhaus et al., 1995). This is useful for general latency studies and for the comparison between groups (Andrà et al., 2009), but no

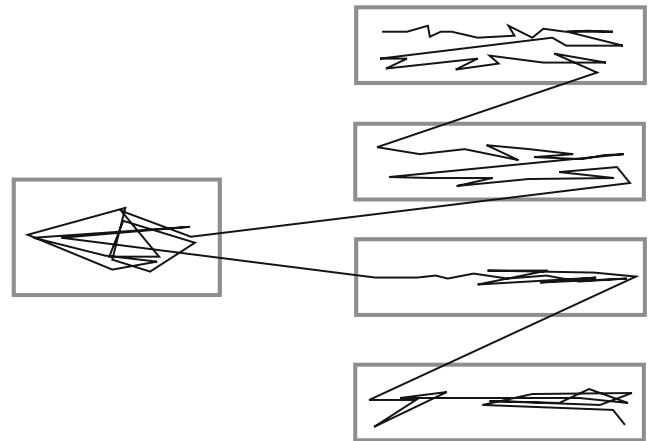


Fig. 2 This fictitious raw scanpath is an overview scan through the five AOIs, but the local versus global, as well as the ambient versus focal, measures would detect much local inspection (short saccades) inside the AOIs, resulting in a high local/focal score

sequence information is used in the measure, so it tells us nothing about the scan sequences of individual participants.

Autocorrelation (ACF) Autocorrelation is defined using a transformation that maps a pattern onto itself, which is why it is called *autocorrelation*. One of the pioneers who applied ACF to eye-movement research is Uttal (1983), who considered discrete 2-D patterns. For a discrete 2-D pattern $f(x, y)$, the autocorrelation function for translations is written as $A(a, b) = \sum_x \sum_y [f(x, y) \cdot f(x + a, y + b)]$. As such, it is a measure of the dynamics over space, while our needs are a measure that encapsulates changes over time, and we will therefore use time as the coordinate (Andrà et al., 2009). Also, autocorrelation does not distinguish between focused and overview behavior.

String edit methods When the scanpaths of each participant are represented with a string of AOIs, the string edit measure can quantify the overall similarity between the scanpaths of two participants—represented as strings of AOIs—by counting the number of edit operations required to transform one of the two scanpaths into the other (Choi et al., 1995; Cristino, Mathôt, Theeuwes, & Gilchrist, 2010). These strings represent sequences over full trials of participants but are used only for pairwise comparisons, not for calculating the frequency of types of sequences. Also, string edit methods do not concern measurements of changes over time.

Transition matrices and Markov chain models A transition matrix is a table of the number of transitions between all pairs of AOIs, exemplified in Table 1, which, for our five AOIs, would be a 5×5 table. As such, it quantifies the frequency of sequences of length 2 AOIs. In our mathematics education project, we would like to study sequences up to length 5, but as we will explain below, transition

Table 1 Transition diagram from the fictitious AOI sequence IAIBACIAIABCDIAIAI. The two most frequent transitions are IA (value 5) and AI (4), while all other transition pairs are infrequent, resulting in 1 s and 0 s in the cells

From	to I	to A	to B	to C	to D
I	–	5	1	0	0
A	4	–	1	1	0
B	0	1	–	1	0
C	1	0	0	–	1
D	1	0	0	0	–

matrices of higher orders become very complex as sequence length increases, because of the exponential growth of the number of different sequences. We have found only one study using transition matrix analysis with sequences of length 3—namely, Ahlstrom and Friedman-Berg (2006), who studied transition sequences in an air traffic controller’s weather station. The authors found very sparse matrices and almost no frequent sequences of length 3. Markov chain models are a variety of transition matrices quantifying the probability of sequences of length 2. When successive transition matrices (“chaining” them) are multiplied, the probability of longer paths can be calculated, but it is not a calculation that can give us the relative prevalence of selected groups of longer sequences (such as overview vs. focused behavior), but only probabilistic estimations of how likely longer single sequences are (Ross, 2006).

Triplet frequencies Groner et al. (1984) counted triplets of AOIs, sequences of length 3, from a face scene with seven AOIs. They found that back-and-forth movements between the eyes were the most common triplets. This is close to our measure, but we need to extend it to longer sequences and add a method for statistical analysis that works with all numbers of AOIs and all sequence lengths.

This overview shows that if we are interested in counting the number of sequences of a particular kind, current measures are limited. None of the measures described can answer questions about how common overall scanning of AOI is, in comparison with movements back and forth between a few AOIs.

Transition sequences with unique AOIs: The first step toward a new measure

For purposes of presentation, this section introduces the new measure by an example using fictitious data. We then show how the measure works on the real data from a large data set that it was developed for. In this part, we

anonymize the AOIs, but later in the article, we again differentiate between all five AOIs.

A transition matrix is a full catalogue of all sequences of length equal to the dimensionality of the matrix. For instance, keeping score of all sequences of length 2 results in a transition matrix with $5^2 - 5 = 20$ cells, as in Table 1. With sequence length 5, when there are n AOIs, a five-dimensional transition matrix representation of all sequences requires $5 \cdot (5 - 1)^{n-2}$ cells, which equals 1,280 cells for five AOIs. Unless enormous amounts of data are recorded, the vast majority of these cells will be empty. This exponential growth and the resulting sparseness of transition matrices as the sequence length increases can be dealt with in two very different ways. The first approach is the probabilistic method of the Markov chains—that is, to convert the transition frequencies to probabilities and ignore probabilities at or close to zero. Since this leaves us only with the possibility of calculating the most probable path at a given state, but not frequencies for all kinds of sequences, Markov chain models are not a viable solution.

The second approach to the exponential growth problem is to group the 1,280 sequences into a smaller number of categories. This gives cells with enough data in each even when there are smaller amounts of data, but the grouping of sequences into categories must be done in a way that makes sense to the research question. Fortunately, this is possible in an analysis of overview versus focused eye-movement behavior: If we have five AOIs (I, A, B, C, and D), we want to know how common it is that participants sequence all five AOIs in a row (represented as ACDBI), as compared with sequences that involve only two AOIs (such as IAIAI or BDBDB). In our mathematics stimulus of Fig. 1, looking at all five AOIs in a row would be indicative of an overview search, while an IAIAI search sequence is likely to involve some form of comparison between the I and A content. In general, a long sequence with only two AOIs involves a more focused search than one where all five AOIs are involved. We will call such sequences *unique-2* and *unique-5*. Using uniqueness analysis allows our analysis to work with four classes of sequence types, rather than with 1,280 cells. Of course, viewing IAIAI and BDBDB as equal is a strong reduction of the amount of information in the data, so that we can no longer differentiate between *which* AOIs the scanpath traverses. We will return to this issue later.

The measure is calculated in the following way. Suppose, for instance, that for a hypothetical participant and trial, we record a scanpath over AOIs. We transform it into an AOI sequence, such as IAIBACIAIABCDIAIAI, where each letter is an entry into an AOI. Consecutive fixations in the same AOI are reduced to one, as in a

compressed string edit representation. String truncation due to length differences is not necessary. TS–AOI sequences are then defined by letting a window of size 5 travel along the AOI sequence. First, the window will encounter IAIBA and see that there are *three unique AOIs* (i.e., I, A, B). Next, we move the window one step and find AIBAC, which has four unique AOIs. We continue like this, until we have reached IAIAI at the end, which has only two unique AOIs. In total, we will have 14 AOI uniqueness numbers from the recorded sequence of 18 AOIs—namely, 3–4–4–4–3–3–3–4–5–5–5–4–3–2. We now count how many unique-2, unique-3, unique-4, and unique-5 AOIs there are in this sequence, and we find one unique-2, five unique-3, five unique-4, and three unique-5AOIs. In Table 2, an example is shown.

Table 2 exemplifies a simple analysis of the sequence of AOI uniqueness numbers. We compare the number of two-TS–AOI windows with the three-, four- and five-unique windows. To our hypothetical participant (number 1), we have added 4 more fictitious participants and can make a table of all 5 with these numbers. These 5 participants exhibit very different behaviors. Participant 2 has many 5s, which means that this participant has more or less circulated all five AOIs round and round. Participant 3 has very many 2s, which corresponds to making many pairwise comparisons, such as IAIAI, although we do not know which pair of AOIs. The last 2 participants have tendencies in each direction: participant 4 toward pairwise comparisons and participant 5 to circling the AOIs.

After this introduction, let us now look at how the measure works with real data, instead of fictitious data. Figure 3 shows the data from all 43 trials and 41 participants divided into the two groups of high and low ability. The tendency is clear: High-ability students make more focused sequences between just a few AOIs (unique-2 and unique-3), while low-ability students make more overview scans with sequences that involve four and five AOIs. Figure 4 shows the first five transitions in all trials—that is, the first window of five AOIs in the sequence from just after the onset of the stimulus. The only difference is that low-ability participants make even more unique-5 and

Table 2 Number of windows with two, three, four, or five unique AOIs for each of 5 fictitious participants

	Unique-2s	Unique-3s	Unique-4s	Unique-5s
Part. 1	1	5	5	3
Part. 2	0	1	1	12
Part. 3	11	2	1	0
Part. 4	9	2	3	0
Part. 5	0	4	4	6

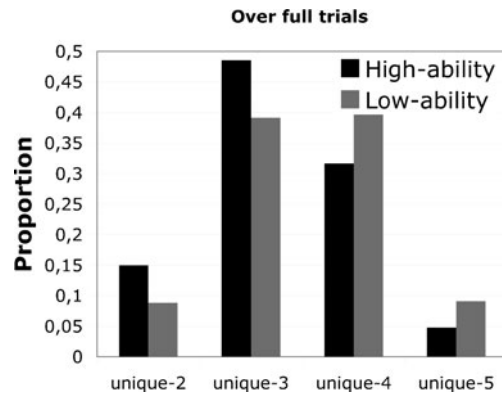


Fig. 3 High-ability students make more focused (unique-2 and unique-3) sequences over the entire trial

fewer unique-2 sequences just after onset. But the really large difference in real data is just before the click to answer, shown in Fig. 5. Toward the very end of trials, high-ability students make 25% pairwise comparisons—that is, cases with five consecutive dwells, in which only two unique AOIs were visited—and another 43% unique-3 sequences.

Our proposed measure is thus able to detect difference in real data, relating to focused versus overview scans across AOIs. In the next section, we will look at how the measure behaves over time and then how we can calculate statistics on the measure.

Transition sequences with unique AOIs: time turns up

In Figs. 4–5, we saw a difference in scanning behavior between our two participant groups at the onset of trials, as compared with the end. We will now look more closely at the behavior over time. For each moment in time—that is, each stop of the window over the AOI sequences in the upper part of Table 3—our measure calculates a histogram, resulting in the lower part of Table 3. When the histograms

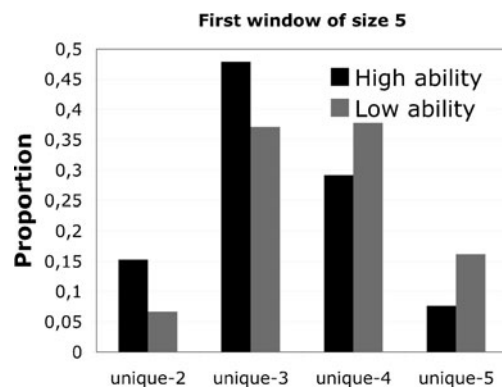


Fig. 4 The first window of five transitions, with only small differences from the data for the whole trial

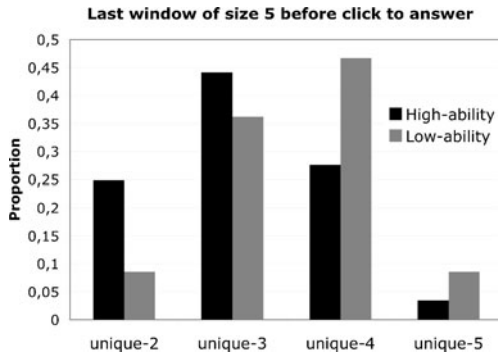


Fig. 5 The very last window before participants clicked the selected alternative and ended the trial

are put into a sequence, we can see and analyze the development over time. Most of the fictitious participants have 4 s and 5 s at the early stages, indicating an early scanning behavior over many AOIs, while during later stages in the table, they generally have lower uniqueness numbers, indicating that they go back and forth between a small number of AOIs. As above, these data reveal sequence information, but not which AOIs are visited.

Let us now return to the real data. In order to simplify graphs, and because they appear to show a very similar effect (according to Fig. 4), we will from now on collapse unique-2 and unique-3 into one group for focused behavior and unique-4 and unique-5 into a single group for overview scanning.² Since focused and overview curves are complementary (they sum to one), we highlight the focused (unique-2 and -3) curve.

Figure 6 shows data over time for a single high-ability participant and all 43 tasks. On average, this participant starts with a quick pairwise looking behavior, followed by an equally short scan over many or all AOIs, and then more in-depth comparisons of fewer AOIs. Figure 7 shows a single low-ability participant with a less distinct pattern.

In Fig. 8, we summarize the data for all trials and all participants, separated into high- and low-ability. We show only the proportion of focused behavior (unique-2 and unique-3 data). The graph shows that high-ability students make more focused movements than do low-ability students throughout the entire trial (upper graph), but the

² Our research question examined two opposite behaviors (comparing two or few items vs. looking at almost everything), which governed the separation between focused behavior and overview scanning. Since the data for unique-2 and unique-3 were similar, as well as the data for unique-4 and unique-5, we decided to simplify the graphs by collapsing them. This was done purely to facilitate reading and the continuing discussion of our new measure in this article. We acknowledge the fact that for some research questions, it may be necessary to include only comparisons between exactly two items, thus including only unique-2 strings, or that for some questions, it may be appropriate to include only overview scanning, where all items are included. In our example, it would refer to including only unique-5.

Table 3 Upper part shows sequences of the number of unique AOIs for successive moving windows. The lower part of the table shows the successive histograms—the four vertical values belonging to each uniqueness number—calculated for the values for all participants in the same column. Time goes from left to right

	Sequences of Uniqueness Values
Participant 1	4-3-4-4-3-3-3-4-5-5-5-4-3-2
Participant 2	5-4-5-4-3-4-5-4-3-2-3-2-3-2
Participant 3	4-4-5-3-3-4-3-3-3-2-3-4-2-3
Participant 4	5-5-5-4-3-2-2-3-3-2-3-2-3-2
Participant 5	5-4-5-5-4-3-3-4-3-3-2-2-3-2
	Histograms Over Time
Number of unique-2s	0-0-0-0-0-1-1-0-0-3-1-3-1-4
Number of unique-3s	0-1-0-1-4-2-3-2-4-1-3-0-4-1
Number of unique-4s	2-3-1-3-1-2-0-3-0-0-0-2-0-0
Number of unique-5s	3-1-4-1-0-0-1-0-1-1-1-0-0-0

difference is larger in the beginning (transition number 5–6, since number 1–4 was consumed by the first window) and after around transition 9. A binomial test, which is used to test whether a value is significantly different from chance level, shows that high-ability participants make significantly more focused sequences than chance throughout. As later calculation will show, the chance probability for the collapsed unique-2 and unique-3 categories is about .3436. The figure shows how the low-ability group hovers around .5 throughout the graph, significantly more focused than chance only at some points in time.

As a consequence, since high-ability participants make significantly more focused sequences than chance throughout the graph, and the low-ability group hovers around .5 throughout the graph (significantly more focused than chance only at some points in time), it is possible to compare the two groups with each other at different time points and conclude that, in most cases, between-group differences are significant. This result can be achieved also

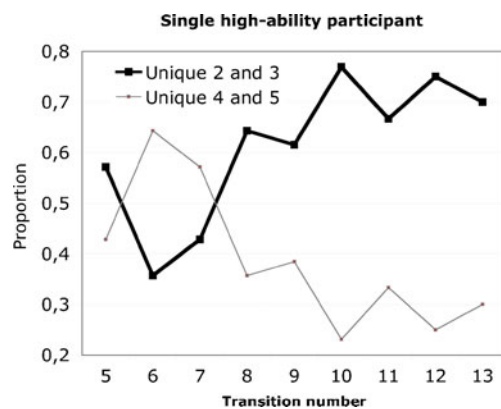


Fig. 6 Behavior of a typical single high-ability participant over time, collapsed over all 43 stimuli. The transition number on the x-axis starts with 5 because transitions 1–5 are included in the first window

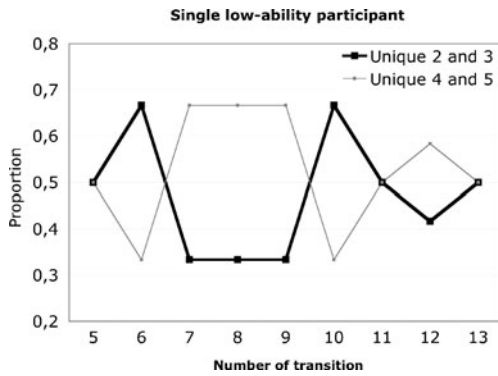


Fig. 7 Data from a typical single low-ability participant, all 43 stimuli

using, for example, a *t*-test for independent populations at each time point, but in this case, the information that high-ability participants are above chance level at each time point would be missed.

Observe that the time dimension in the figures signifies the order of the windows for calculating the number of unique AOIs. Since the first window always consumes four transitions, time in these diagrams will start at the fifth transition, and its unit will be number of transitions from onset. The first value in the curve therefore represents not just a point in time, but a whole window that can be a few seconds long.

Furthermore, this over-time measure ignores whether a participant dwelled for a longer or shorter time at a particular AOI between transitions. It reflects only transition order. When we summarize data from the same transition number across several participants, we must be aware that these data may originate from quite different points in actual time.

A simple test to verify that this is not the case would be to examine whether the number of transitions per time unit is the same between participants, which can be assessed using an ANOVA test. However, many times, in practice,

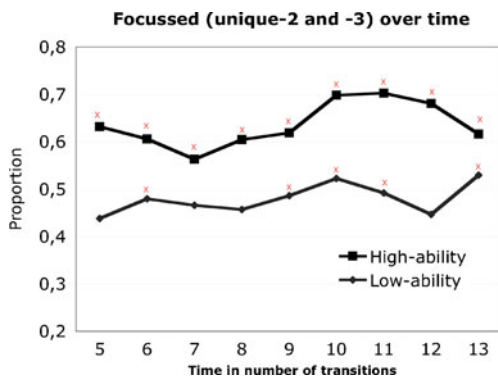


Fig. 8 All participants, by groups, and all 43 stimuli. Xs indicate that the value is significantly different from chance level (34%), according to a binomial test

this simple test is useless, since dwell times within AOIs often differ significantly between participants.

Hence, it could be useful to test whether dwell times can be considered as exponentially distributed (Ross, 2006). If dwell times have an exponential distribution, in fact, the assumptions for applying an ANOVA to our measure hold, and this is a classic result in literature (see, e.g., Ross, 2006). Testing the exponential distribution of dwell times can be performed using the classic chi-square test or the Kolmogorov–Smirnov test (Ross, 2006) on the distribution of dwell times.

Since dwell times often have an exponential distribution (Holmqvist et al., 2011), our method does not lose generality because of this constraint. To sum up, even if dwell times differ significantly among participants (as an ANOVA test may highlight), it is enough for applying this measure that dwell times have an exponential distribution. This condition is satisfied in most experimental situations.

The participants in the study that we use for exemplification ended their trials by selecting the alternative they judged was the correct one. If we want to analyze their behavior just before the click, when they were ready to give their answer, we have to align the sequences at their endpoints of the trials, before we summarize over participants. Figure 9 shows such data. We can clearly see that the high-ability participants stick with their focused eye-movement behavior to the very end.

Two settings for the transition sequences with unique AOIs

There are two settings for this TS–AOI measure. The first is its *window size*, which we call *w*. All the examples above use a five-AOI window; that is, the window stretches over five AOIs in the recorded AOI sequence, because there are

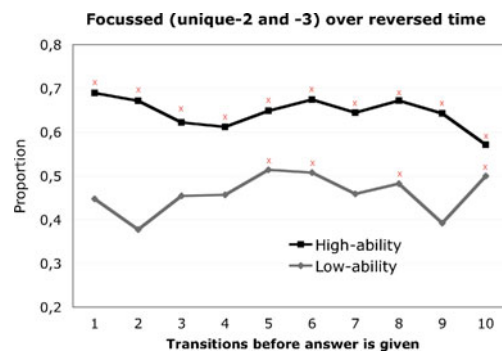


Fig. 9 These data are aligned so that the end of trial corresponds to time 0 in the diagram. Time 1 is the last window of five AOIs before the end of trial, time 2 the second to last, and so forth. This also means that the data are reversed, so time is ordered in the opposite direction from that in the other graphs. Xs indicate that the value is significantly different from chance level (34%), according to a binomial test

five AOIs in the stimulus and it is possible to visit all five in one consecutive sweep. Since five AOIs is the maximum number of AOIs to visit, there is no need to have a larger window than there are AOIs.

But what is the maximal window size? In a stimulus with 10 AOIs, is an AOI window size of 10 reasonable? There are at least two arguments that can be made. First, the statistical treatment of this type of data indicates that 10 AOIs and a window size of 10 are feasible, but with more AOIs and longer sequences, data become more and more skewed to the higher uniqueness categories. Second, it would be possible to argue that in cases in which low uniqueness values are taken to indicate focused comparisons, the window size should probably not be larger than the number of items that a participant can hold in memory (Miller, 1956), because otherwise we would not be measuring comparisons. This is likely to be in the same general range, not much more than 10.

The other setting is the *step size* when the window is moved ahead in the recorded AOI sequence. We have used a one-step setting above. The shortest possible step size gives the best temporal resolution to the measure, but a moving window undermines the independency requirement for many statistical tests. One solution would be to move the window not one step, but five steps, so as to get successive samples of size 5. This will essentially give independence (since no data points in one window are present in the adjacent window) and allow for classical hypothesis testing, but at the cost of less data. Fortunately, other statistical solutions exist.

Transition sequences with unique AOIs: statistical methods

In this section, we present statistical methods for the measure we propose. As was stated earlier, we wanted to know what students are doing while engaged in a mathematical task and to find out what kind of differences there are in students' eye-movement behavior over time. In the following, we will exemplify how to (1) compare transition sequences with unique AOIs over time, (2) compare different intervals (i.e., eye-movement behavior early vs. late in a task), and (3) perform statistical testing.

We first address the independence issue: When a moving window is used over the original sequence of transitions, values in the sequence of uniqueness values are not independent of their neighboring values. This undermines some uses of variance tests such as ANOVAs and the *t*-test. In particular, we cannot test in Fig. 8 whether the proportion of focused (unique-2 and unique-3) sequences at transition 5 is significantly lower than that at transition 6, because data at transitions 5 and 6 are calculated from

overlapping windows and, hence, share the same origin. Nor can we use these well-known classical tests, such as ANOVAs, to compare the levels of our groups over time.

However, there is an alternative statistical method for the same type of comparison, based on so-called time series analysis (Box, Jenkins, & Reinsel, 1994), that can be carried out in most statistical softwares (SPSS, R, etc.). It resembles work by Uttal (1975) but uses the temporal coordinate, *t*, to compute the correlation between a certain instant *t* and another instant *t + l*, where *l* is the lag. The following step-by-step description is a summary of the procedure, but consult literature on time series analysis for details.

Assume that we have two series of uniqueness proportions, as in Figs. 8 and 9, that we wish to compare between groups or conditions or to compare an earlier stage against a later. Comparing the high-ability participants against the low-ability participants therefore corresponds to performing this test for *both* series of values and making sure that they both pass the test (insignificant autocorrelation values and normal distribution) and that they are centered around different averages (for instance, 0.7 vs. 0.5).

1. For each of the two series—for instance, for the high-ability participants in Fig. 8—investigate whether the curve is approximately constant at a certain level (such as 0.7). If it is not, derivate it to achieve a constant level.
2. Calculate the averages of both series (now approximately constant). It is these that we actually compare.
3. Subtract the average of each series from the values in it, so as to make the series centered around zero. Check that the centered values have a normal distribution with a mean of 0.
4. Estimate what are known as the autocorrelation functions (ACFs) and the partial autocorrelation functions (PACFs) on the centered values. The ACF and the PACF values are insignificant—that is, approximately zero³—for a lag greater than a (small) value. This means that the values are not correlated with each other for a long time.⁴ To give an idea, this can be considered as a further check of constantness in the series.
5. Estimate the ACF and the PACF on the residuals. The two series are significantly zero if the residuals are randomly distributed around zero. Such a test does not give information on the behavior, but it simply allows us to verify a posteriori the goodness-of-fit of the model to data. In the case in which residuals do not

³ When an appropriate statistical software is used, confidence intervals for the zero values of the ACF and the PACF are provided by it.

⁴ Assuming that the ACF and the PACF are approximately zero for a lag greater than a certain value does not lead to a loss of generality of the model. It sounds reasonable, in fact, to suppose that, after a certain time interval, values would not be autocorrelated.

satisfy this condition, it could be necessary to change the parameters that characterize the model.

The second statistical issue that needs to be addressed is the *chance probabilities* of the distribution of data across the four classes with uniqueness values of 2–5, which we need for binomial tests. It may be tempting to assume that the four classes (unique-2 to unique-5) would exhibit an equal distribution of 25% chance level to each. This is not the case, however.

In order to correctly calculate chance levels, we employ MATLAB calculations and combinatorics. It turns out that the actual distributions of the number of possible unique- k sequences depends heavily on the number of AOIs n , the value k , and, to some extent, the window size w . Table 4 shows the distribution of all possible strings for a window length w of five and the number of AOIs n from 2 to 12. In our mathematic-problem-solving study, we used $w = 5$ and $n = 5$, which means that the chance probability of making a unique-2 sequence is 1/21 of the chance for a unique-3 sequence, and even a sixth compared with making a unique-5 sequence. Recalculated as normalized probabilities, we have (u-2,.0156; u-3,.328; u-4,.562; u-5,.0938). This chance distribution resembles that of the low-ability participants just before they clicked to answer (Fig. 5), but even then, these participants make many more u-2 sequences (8%) than they would have if they had looked at the stimulus completely at random (1.5%). After we have calculated these chance levels, the true binomial test allows us to correctly test the hypothesis that, for instance, the high-ability students make many more unique-2 transition sequences than chance.

Table 4 gives the basis for calculating chance probabilities for many types of studies where the window size $w = 5$.

Table 4 The number of possible unique-2 (u-2) to unique-5 (u-5) sequences given a number n of AOIs and a window length $w = 5$. “All” is the total number of transition sequences of length w in an n -dimensional transition matrix, AOI repetitions excluded

w	n	u-2	u-3	u-4	u-5	All
5	2	20				20
5	3	20	140			160
5	4	20	280	240		540
5	5	20	420	720	120	1,280
5	6	20	560	1,440	480	2,500
5	7	20	700	2,400	1,200	4,320
5	8	20	840	3,600	2,400	6,860
5	9	20	980	5,040	4,200	10,240
5	10	20	1,120	6,720	6,720	14,580
5	11	20	1,260	8,640	10,080	20,000
5	12	20	1,400	10,800	14,400	26,620

But if the number of AOIs n is larger than 12, it is useful to know that the numbers in Table 4 are calculated from Eqs. 1–5 below. Note that these are valid only for $w = 5$.

$$u - 2 = w \cdot (w - 1) \quad (1)$$

$$u - 3 = w \cdot (w - 1) \cdot (n - 2)(2^{(w-2)} - 1) \quad (2)$$

$$u - 4 = 6 \cdot w \cdot (w - 1) \cdot ((n - 2)^2 - (n - 2)) \quad (3)$$

$$u - 5 = w \cdot (w - 1) \cdot ((n - 3)^3 - (n - 3)) \quad (4)$$

$$\text{All} = w \cdot (w - 1)(n - 1)^3 \quad (5)$$

Rather than presenting tables with chance-level calculations for other w s with large numbers, we present the equations for calculating chance levels for window sizes 4 and 6 and for all n . For window sizes $w = 4$, the corresponding Eqs. 6–9 give the number of possible transition sequences for all n s. Note that there are only small differences from Eqs. 1–5; in particular, Eq. 9 is quadratic, and Eq. 5 is cubic. Also, compare Eqs. 4 and 8. Nevertheless, the calculation for unique-2 and unique-3 values is the same for all examined window sizes:

$$u - 2 = w \cdot (w - 1) \quad (6)$$

$$u - 3 = w \cdot (w - 1) \cdot (n - 2)(2^{(w-2)} - 1) \quad (7)$$

$$u - 4 = w \cdot (w - 1) \cdot ((n - 2)^2 - (n - 2)) \quad (8)$$

$$\text{All} = w \cdot (w - 1)(n - 1)^2 \quad (9)$$

For $w = 6$, the equations for calculating all possible transition sequences change some more, as shown in Eqs. 10–15. For window size $w = 7$, the equations are quite complex, but one thing remains constant: As we increase the window size—which corresponds to increasing the dimensionality of the underlying transition matrix—the number of cases with low uniqueness values remains slow-growing, and the bulk of the growth in the transition matrix is consumed by the higher uniqueness categories. In fact, the number of unique-2 sequences is always constant in n , the number of unique-3 always linear in n , and generally it

appears that the number of possible unique- k transition sequences is described by a polynomial of grade $k-2$:

$$u - 2 = w \cdot (w - 1) \quad (10)$$

$$u - 3 = w \cdot (w - 1) \cdot (n - 2)(2^{(w-2)} - 1) \quad (11)$$

$$u - 4 = 25 \cdot w \cdot (w - 1) \cdot ((n - 2)^2 - (n - 2)) \quad (12)$$

$$u - 5 = 10 \cdot w \cdot (w - 1) \cdot ((n - 3)^3 - (n - 3)) \quad (13)$$

$$u - 6 = w \cdot (w - 1) \cdot ((n - 2)^4 - 6 \cdot (n - 2)^3 + 11 \cdot (n - 2)^2 - 6 \cdot (n - 2)) \quad (14)$$

$$\text{All} = w \cdot (w - 1)(n - 1)^4 \quad (15)$$

In other words, the more AOIs we have, and the longer the sequences are that we measure, the less likely it is that our participant is going to make pairwise unique-2 movements back and forth between two AOIs. For instance, with a window of $w = 7$ AOIs and a total of $n = 11$ AOIs in the stimulus, the unique-6 sequences are 45,360 times more likely, as compared with unique-2 sequences.

Transition sequences with unique AOIs: When the information in a specific AOI counts

One major drawback with the method above is that we disregard the AOI identity completely. For instance, the measure counts how many unique-2 sequences there were, but it cannot distinguish between AIAIA and BCBCB, although these two sequences appear to correspond to two very different cognitive processes. In this section, we present a development of the measure that includes AOI identity.

Let us continue using the number of AOIs $n = 5$ and window size $w = 5$, as above. We will now assume that AIAIA and IAIAI can be considered identical but that they are different from BCBCB, as in Fig. 10. More generally, sequences with *the same uniqueness number and the same AOIs* will be considered identical, even if the order is different or the number of instances of each is different. For instance, the two unique-4 sequences IACBI and CIABC are considered equal, but they are different from DBAIA.

Statistically, this is just a further division of each uniqueness category into subcategories. The unique-2 category of sequences have $\binom{5}{2} = 10$ such subcategories, one containing the two AIAIA and IAIAI, and one

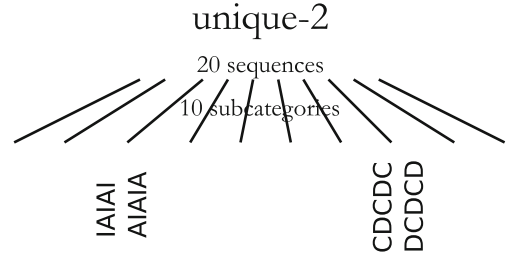


Fig. 10 Of the total 20 unique-2 sequences ($n = 5$, $w = 5$), there are 10 subcategories, each with two members. The two subcategories shown are the unique-2 with I and A subcategory of sequences and the unique-2 with C and D

subcategory each for the corresponding sequences involving the AOI pairs (I,B), (I,C), (I,D), (A,B), (A,C), (A,D), (B,C), (B,D), and (C,D). Each such category has two members, which are sequences.

For the unique-3 category, there are $\binom{5}{3} = 10$ subcategories, although now each subcategory involves three AOIs, as in Fig. 11. Similarly, the unique-4 category has five—that is, $\binom{5}{4}$ —subcategories with four identical AOIs in each. The unique-5 category does not have any subcategories, since $\binom{5}{5} = 1$.

In total, we have 26 subcategories, rather than four categories, when the number of AOIs $n = 5$ and window size $w = 5$. The increase in number of categories allows finer experimental distinctions to be made but will also require more data to distribute among the categories, or significance levels will not be reached. In general, the number of categories can be calculated as in Eq. 16.

$$\text{Number of subcategories} = \sum_{i=2}^{\min(n,w-1)} \binom{n}{i} \quad (16)$$

Baseline chance probabilities for each subcategory can easily be calculated by dividing the values in Table 4 by the corresponding number of subcategories. For instance, we read a total of 420 possible unique-3 sequences. Divided by the $\binom{5}{3} = 10$ subcategories, we have a chance level of each subcategory of 42 possible sequences, or a normalized base probability of $42/1280 = 3.2\%$ for each unique-3 subcategory. The other base line probabilities for $n = 5$ and

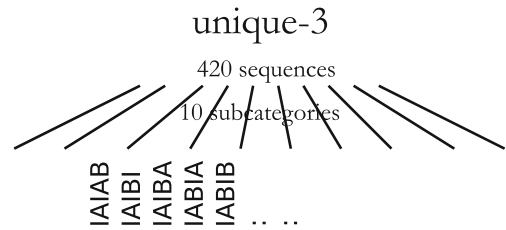


Fig. 11 Of the total 420 unique-3 sequences ($n = 5$, $w = 5$), there are 10 subcategories, each with 42 members. The single subcategory partially shown is the unique-3 subcategory with I, A, and B

$w = 5$ range from 0.15% ($\frac{2}{1280}$) each for the two unique-2 subcategories to 11.25% ($\frac{144}{1280}$) for the five unique-4 subcategories.

Discussion

We have proposed the TS–AOI measure to sort sequences of transitions into meaningful groups. None of the existing measures we mentioned at the outset of this article combine AOI-based sequences with an analysis of overview and focused visual behavior. Although they both adopt similar AOI sequences, the TS–AOI measure differs from the string edit method by analyzing frequencies, rather than pairwise similarities. The intended goal of the TS–AOI measure is similar to that of the local versus global and the ambient versus focal measures—namely, to classify focused versus overview looking—but the TS–AOI measure works on AOI sequence data, which allows for detailed classification of sequences founded on a meaningful division of space.

In its basic form, the TS–AOI measure has four groups when the sequence length is five and, more generally, $\min(n-1, w-1)$ groups when n is the number of AOIs and w is the sequence length. This is a reduction in the number of groups in data from the exponential number $w \cdot (w-1)^{n-2}$ of cells in the w -dimensional transition matrix for n AOIs to a number ($\min(n-1, w-1)$) that is almost constant. Still, meaningful analyses can be made, as we have shown with the mathematical-problem-solving data.

In its second form, the TS–AOI measure distinguishes sequences in each uniqueness category on the basis of the AOIs in it. With length 5 and five AOIs, there are 26 such groups of sequences. When n grows, the number of groups of sequences grows exponentially, but much more slowly than the number of cells in the corresponding transition matrix.

It is easy to write a program to calculate uniqueness values, and computational tractability is very good, with a constant number of calculations per sequence. We have given examples of methods that allow for statistical analysis of the output from the measures.

Using the measure on real data, we have shown significant differences between high- and low-ability participants in our mathematical-problem-solving task. The high-ability group makes more focused sequences throughout the trials, and this difference reaches its peak just before the click is made to answer. Furthermore, when applying the measure to data from mathematical problem solving, we have shown a strong task- and competence-related effect on scanning: High-ability participants predominantly make focused (unique-3 and unique-2) transition sequences at the onset of the stimulus. This viewing behavior differs from a long line of research on picture viewing

(Buswell, 1935; Unema et al., 2005), which has shown that overview scanning starts very soon after picture onset.⁵

The possibility of making over-time analysis of uniqueness values is another attractive feature of the measure. Although the temporal resolution is coarse, it is sufficient for applied studies like ours. This problem stems from using AOI sequences and is general for all measures operating on AOI sequences.

A clear limitation is the upper limit on the number of AOIs that can be included. In our example, we used 5 AOIs. Using the TS–AOI measure with 100 AOIs would not be successful, simply because with such a large n , the chance baseline will be totally dominated by unique-5 sequences with $w = 5$. Our estimations indicate that the measure works for about 3–12 AOIs. Nevertheless, this range covers most typical AOI numbers in studies. Sequence lengths could be between 3 and 10 AOIs. Depending on the research question and how semantic meaning is defined, the method can be used with stimuli of higher complexity, including paintings or real-world images—for example, examining the skills of a person processing complex stimuli requiring reading and/or integration of various information, such as weather forecasting, medical diagnosis of x-ray images, and so forth.

A further limitation is that dwell time in AOIs—that is, the duration between transitions—is not included in the proposed measure. Low dwell times mean a short inspection time in the AOIs and would indicate different processing, as compared with long dwells. Dwell time can be additionally analyzed in a Markov chain model when the time is assumed to be exponentially distributed (Ross, 2006).

All in all, we consider the unique number of AOIs measure to be a useful and productive addition to previous measures of scanning across multiple AOIs.

References

- Ahlstrom, U., & Friedman-Berg, F. (2006). *Controller scan-path behavior during severe weather avoidance* (Tech.Rep.DOT/FAA/TC-06/07). William J. Hughes Technical Center, Atlantic City International Airport, NJ 08405 Federal Aviation Administration; NAS Human Factors Group, AJP-7132.
- Andrà, C., Arzarello, F., Ferrara, F., Holmqvist, K., Lindstrom, P., Robutti, O., & Sabena, Cristina (2009). How students read mathematical representations: An eye tracking study. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Eds.), *Proceeding of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 49–56). Thessaloniki, Greece.

⁵ Our mathematical stimuli are not pictures in the sense of a painting. However, we think that they are complex and that the contrasting result with picture viewing is interesting to highlight. From a pedagogical perspective, differences in viewing behavior is interesting because it can be informative with respect to the progress of a student's learning.

- Box, G. E. P., Jenkins, G. M., & Reinsel, G. (1994). *Time series analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- Buswell, G. (1935). *How people look at pictures*. Chicago: University of Chicago Press.
- Choi, Y. S., Mosley, A. D., & Stark, L. W. (1995). "Starkfest" vision and clinic science (Special issue: String editing analysis of human visual search). *Optometry and Vision Science*, 72, 439–451.
- Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42, 692–700.
- Groner, R., Walder, F., & Groner, M. (1984). Looking at faces: Local and global aspects of scanpaths. In A. G. Gale & F. Johnson (Eds.), *Theoretical and applied aspects of eye movement research* (pp. 523–533). Amsterdam: North-Holland.
- Holmqvist, K., Nyström, N., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.
- Mannan, S., Ruddock, K., & Wooding, D. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10, 165–188.
- Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity to process information. *Psychology Review*, 63, 81–97.
- Ponsoda, V., Scott, D., & Findlay, J. M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica*, 88, 167–185.
- Ross, S. (2006). *Introduction to probability models*. Amsterdam: Academic Press.
- Simola, J., Salojärvi, J., & Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9, 237–251.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45, 643–659.
- Unema, P., Pannasch, S., Joos, M., & Velichkovsky, B. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12, 473–494.
- Uttal, W. R. (1975). *An autocorrelation theory of form detection*. Hillsdale, NJ: Erlbaum.
- Uttal, W. R. (1983). *Visual form detection in 3-dimensional space*. Hillsdale, NJ: Erlbaum.
- Zangemeister, W. H., Sherman, K., & Stark, L. (1995). Evidence for a global scanpath strategy in viewing abstract compared with realistic images. *Neuropsychologia*, 33, 1009–1025.