Looking with the (computer) mouse: How to unveil problem-solving strategies in matrix reasoning without eye-tracking

Guillaume Rivollier, Jean-Charles Quinton, Corentin Gonthier, Annique Smeding

Looking with the (computer) mouse:

How to unveil problem-solving strategies in matrix reasoning without eye-tracking

Guillaume Rivollier [1, 3], Jean-Charles Quinton [1], Corentin Gonthier [2], & Annique Smeding [3]

Author Note

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2] Université Rennes 2, LP3C EA 1285, 35000 Rennes, France
[3] Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, LIP/PC2S, 38000 Grenoble, France

This work was supported by the Pôle Grenoble Cognition, and the French National Research Agency in the framework of the "Investissements d'avenir" programs ANR-15-IDEX-02 and ANR-11-LABX-0025-01.
We would like to thank Alisée Bruno for her help with data collection in Experiment 1.

*Corresponding authors: Annique Smeding, BP 1104, 73011 Chambéry cedex, France. Phone: +33 4 79 75 85 89; e-mail: annique.smeding@univ-smb.fr
Jean-Charles Quinton, LJK - Bâtiment IMAG, 700 Avenue Centrale, 38401 Domaine Universitaire de Saint-Martin-d'Hères, Phone: +33 4 57 42 17 78, e-mail: quintonj@univ-grenoble-alpes.fr

Abstract

Problem-solving strategies in visual reasoning tasks are often studied based on the analysis of eye movements, which yields high-quality data, but is costly and difficult to implement on a large scale. We devised a new graphical user interface for matrix reasoning tasks, where the analysis of computer mouse movements makes it possible to investigate item exploration and, in turn, problem-solving strategies. While relying on the same active perception principles underlying eye-tracking (ET) research, this approach has the additional advantages of being user-friendly and easy to implement in real-world testing conditions, and records only voluntary decisions. A pilot study confirmed that embedding items of Raven's Advanced Progressive Matrices (APM) in the interface did not significantly alter its psychometric properties. Experiment 1 indicated that mouse-based exploration indices, when used to assess two major problem-solving strategies in the APM, are related to final performance – as has been found in past ET research. Experiment 2 suggested that constraining some features of the interface favored the adoption of the more efficient solving strategy for some participants. Overall, the findings support the relevance of the present methodology for accessing and manipulating problem-solving strategies.

Keywords: Active perception; Matrix reasoning; Strategy use; Graphical user interface; Computer mouse movements

Looking with the (computer) mouse: How to unveil problem-solving strategies in matrix reasoning without eye-tracking

Human intelligence is most often assessed via performance on reasoning tasks: performance tests wherein participants have to solve problems through inductive or deductive reasoning. Matrix-like reasoning tests are a prototypical example of a reasoning task (Carpenter, Just, & Shell, 1990), and have long been used in intelligence tests (Wechsler, 2008). In particular, Raven's Advanced Progressive Matrices (APM; Raven & Court, 1998b) are often used by investigators. This type of task requires participants to find the missing piece of a visual display among several possibilities. Each APM item comprises two parts (see Figure 1, which also outlines the setup for the present study). The top part is a 3 x 3 matrix; patterns are visible in eight out of nine cells, while the bottom right cell is always empty. The objective is to find the missing pattern, which obeys a variety of logical rules (Carpenter et al., 1990; Vigneau & Bors, 2008). The bottom part contains eight response alternatives, of which only one correctly completes the matrix.

Matrix reasoning tasks have been used in countless studies and provide useful insights into individual differences in fluid intelligence and their determinants (e.g. Ackerman, Beier, & Boyle, 2005). However, total performance on the APM does not directly inform us about the actual response processes, such as problem-solving strategies, through which participants reach an answer. Our understanding of these processes is still limited, in part because complex and constrained methodologies such as eye-tracking are necessary to capture them (e.g., Carpenter et al., 1990; Vigneau, Caissie, & Bors, 2006).

In the present research, we aimed at going beyond eye-tracking as the default methodology for understanding real-time problem-solving processes on matrix-like intelligence tests. To do so, we devised a new methodology based on computer mouse

movements, while relying on the same active perception principles that underlie eye-tracking research. This approach is easy to implement in real-world testing conditions, preserves the core features of the APM, and produces high quality data as only the voluntary actions of participants are recorded.

Problem-Solving Processes in Raven's Matrices

To understand processes driving performance and individual differences in matrix reasoning tasks, most early work relied on eye-tracking data. For example, the results of a seminal study analyzing eye movements suggested that participants proceed incrementally in the APM: items are decomposed into a series of steps solved sequentially, with subjects performing a series of pairwise comparisons to induce the first rule, then the second rule, and so on (Carpenter et al., 1990). This conclusion was influential in understanding the workings of reasoning in visual tasks, especially its relation with goal management abilities (Carpenter et al., 1990).

Most research on response processes in Raven's matrices has been concerned with the different strategies that can be used to reach an answer. Eye-tracking studies, often combined with concurrent verbal reports, have identified two prototypical strategies (Bethell-Fox, Lohman, & Snow, 1984; Snow, 1978; Snow, 1980). The first strategy, constructive matching, is used when one tries to infer logical rules by focusing on the matrix, mentally reconstructing the missing pattern through deduction, and only then selecting the matching pattern among the response alternatives. The second strategy, response elimination, is used when one compares features from the visible patterns in the matrix part with those of response alternatives, so as to eliminate as many incorrect alternatives as possible.

Past research has shown that studying these two strategies is critical to understanding performance in the task, and especially individual differences in performance. Constructive matching is preferentially used by high-performing individuals, whereas response elimination

is generally used by lower-performing individuals, especially on difficult items (Bethell-Fox, Lohman, & Snow, 1984; Snow, 1978, 1980). In this vein, participants with high working memory capacity, who reliably demonstrate higher performance on Raven's matrices (e.g. Ackerman et al., 2005), have been shown to make greater use of the more complex yet efficient constructive-matching strategy, and to rely less on the simpler and less efficient elimination strategy (Gonthier & Thomassin, 2015; Gonthier & Roulin, 2019; see also Jarosz & Wiley, 2012; Wiley & Jarosz, 2012). Likewise, participants tend to switch from constructive matching to response elimination as items become more difficult, in tandem with a drop of performance (Bethell-Fox et al., 1984; Gonthier & Roulin, 2019).

While early studies based on eye-tracking were mostly interested in identifying possible strategies (Bethell-Fox, Lohman, & Snow, 1984; Snow, 1978; Snow, 1980), one study in particular examined how eye-tracking could be used to extract indices of strategy use at participant and item levels, which could then be related to total APM performance (Vigneau et al., 2006; see also Hayes, Petrov, & Sederberg, 2011). Vigneau and colleagues reasoned that constructive matching and response elimination strategies should be apparent in visual explorations: processing fine-grained visual details defining the pattern features is required for both inferring rules from the matrix part and eliminating response alternatives, which implies that fixation time on a given part of the display is an index of the extent to which the participant analyzes that part of the item. On an a priori basis, the authors therefore assumed that a higher proportion of time spent on the matrix, both in total and prior to looking at possible responses, should reflect a constructive matching strategy. Conversely, less time spent on the matrix and more time on the response alternatives should reflect response elimination. Alternating more often between the matrix and the response alternatives should also reflect response elimination.

Vigneau and colleagues (2006) obtained results quite consistent with their expectations

at the individual level (see Table 1 for a summary of results for their correlation analyses), although results at the item level were less conclusive. Proportional time spent on matrix and latency before the first alternation were positively linked to final performance, whereas proportional time spent on alternatives and the number and rate of alternations were negatively related to final performance. Visual exploration patterns, as evidenced in eye-tracking data, were thus demonstrated to be useful indices to track real-time individual differences in strategy use in the APM. These indices rely on rather coarse information about item exploration, simply opposing matrix and response alternatives. Given the structure and visual presentation of APM items, analyzing fine-grained dynamics of eye-movements beyond fixations on areas of interest is usually not needed to infer resolution strategies. Precision below cell scale holds little informative value, since pattern-defining features spatially overlap within cells.

Therefore, there may be easier ways to capture response strategies in larger scale studies, for instance in online studies where eye-tracking validity is difficult to assess and guarantee. Although there has been a recent surge in low cost eye-tracking devices designed for human-machine interaction and gaming, and such devices are now marginally used for research, they impose heavy constraints and provide limited guarantees regarding data quality (e.g., see Gibaldi, Vanegas, Bex, & Maiello, 2017 for a review of the Tobii EyeX capabilities for research). Of importance here, some fundamental assumptions and constraints still threaten the validity of strategy resolution measurement

First, participants are aware of the apparatus, and cannot simply ignore the fact that their eye-movements are being recorded. The importance of non-consciousness from the participants' perspective varies across research areas, but this is often desired in fields such as social cognition (e.g., research on stereotype threat in reasoning tasks; e.g., Brown & Day, 2006; Régner, Smeding, Gimmig, Thinus-Blanc, Monteil, & Huguet, 2010), and may affect

which strategies participants elect to use. Second, and although calibration procedures are progressively simplified and temporally reduced, they not only make the recording of eye-movements more salient, but also impose an overhead on user interactions. With the perspective of deploying matrix-like reasoning tasks in online studies or under the form of a serious game, reducing casual game onboarding (below 60 seconds) is critical to user retention (Clutch, 2017). Third, partly arbitrary decisions must be made regarding the boundaries of areas of interest, which depend on how much peripheral information is supposed to be processed by the user given the constraints of the task to be performed. The same is true for the algorithms used to distinguish between oculomotor events (including saccades and fixations), whose performance of course depends on the sampling rate or noise, but also on the expected dynamics of eye movements, which in turn depend on the task (Zemblys, Niehorster, Komogortsev, & Holmqvist, 2018). Finally, eye-tracking devices record all eye movements, regardless of their relevance for the problem-solving process at hand. In tasks like the APM, this over-recording adds noise to the data (e.g. Hayes et al., 2011): for example, participants may fixate a given area that is not currently relevant for problem solving, while simultaneously processing relevant information from the peripheral visual field (e.g., through top-down selective attention mechanisms reviewed in Gazzaley & Nobre, 2012), or even while mentally combining previously-encoded information. Another source of noise in eye-tracking data comes from the lack of controllability of eye movements, which can make a fine-grained interpretation of data problematic. For example, the gaze may be directed towards a salient visual cue, even though the cue is not used for problem-solving.

In sum, eye-tracking is a highly useful methodology, but one that is cumbersome in many situations, and whose results suffer from a few biases. The purpose of the present research was to devise a new methodology that could overcome most of these issues.

Unraveling Strategies in Matrix Reasoning based on Mouse Interactions

Eye-tracking can be used for the real-time study of strategies on APM items because respondents simply do not scan, memorize and internally process all information about an item at the same time, given the complexity of the available (visual) information. With the hardest APM items involving up to 5 rules and 8 possible responses integrating many features (see Carpenter et al., 1990), it is necessary to perform eye movements to selectively access visual features, and iteratively infer or test rules. However, other effectors may be used to select relevant features for processing: in the present study, we focused on hand-initiated computer mouse interactions.

To this end, we chose to embed the APM in a graphical user interface where item exploration is achieved, in part, through voluntary computer mouse interactions. The rationale was to keep parts of the item hidden, with the participant having to click with the computer mouse on a given part to display it, much like they would make an eye movement towards this location to uncover the corresponding information. Because our objective was to develop a protocol that made it possible to assess the use of constructive matching and response elimination strategies, we used a simple design decomposing each item into the matrix and the response alternatives.

The rationale of studying response processes through participant explorative movements can be framed in an active perception approach of reasoning tasks. A movement of the mouse aiming to selectively access part of an APM item can be viewed as an epistemic action (Kirsh and Maglio, 1994): an action performed to uncover information that is hidden or hard to compute mentally (by contrast with pragmatic actions, which are performed to bring one physically closer to a goal). This distinction has been generalized in human-machine interaction research (Ware, 2012), as many actions taken on computer interfaces indeed serve the purpose of accessing new information or changing the way it is represented to offload our

cognitive efforts. Likewise, the sensorimotor theory of perceptual consciousness (O'Regan, 2011; O'Regan & Noë, 2001) defines vision not as exploiting the retinal signals and eye muscles, but as the sensorimotor laws underlying visual interactions and their intrinsic properties (e.g., inferring space dimensionality; Philipona, O'regan, Nadal, & Coenen, 2004).

In this sense, eye movements and mouse interactions, despite being associated with different effectors (eye and hand) and therefore motor costs, share much similarity. Since the theorization of motor equivalence by Bernstein (1967), empirical studies have demonstrated how the many degrees of freedom of the human body can be flexibly selected and combined to achieve the same goal, and how part of the dynamics and underlying neural processes may be shared (Kelso et al., 1998). Perception and motor control have been further unified under the free energy principle, also described as active inference in neuroscience (Adams, Shipp, & Friston, 2013), as the two sides of the same underlying surprise minimization mechanism. While perception consists in inference and decisions taken to reduce prediction error on sensory signals, action does the same on proprioceptive signals.

In other words, the inference processes on APM need not be reflected only in the sequence of eye saccades and fixations: it may well be that the rules themselves are represented and embodied as sequences of predictions (e.g., for the constant in a row rule: fixation on a feature, saccade to the right, find the same feature). This form of representation would not be sensor or actuator specific, but instead could be defined more abstractly by regularities or contingencies. Thus, as is the case in examples of sensory substitution (e.g., vision to tactile as early as Bach-y-Rita, Collins, Saunders, White, & Scadden, 1969), we can expect eye-tracking and mouse-tracking-based measures to capture the dynamics of the same processes as long as the underlying sensorimotor laws are maintained (i.e. moving the eyes and/or the mouse over a given part of an item reveals the same set of features).

In line with this view, mouse-tracking has been extensively used to examine dynamic

competition in forced-choice categorization tasks (e.g., Freeman & Ambady, 2010), and has recently become a favored technique to examine real-time decision-making processes underlying categorization in various research areas, including psycholinguistics (e.g., Crossley, Duran, Kim, Lester, & Clark, 2018; Magnuson, 2005; M. J. Spivey, Grosjean, & Knoblich, 2005; Michael J. Spivey & Dale, 2006), social cognition (e.g., Freeman, Pauker, & Sanchez, 2016; Smeding, Quinton, Lauer, Barca, & Pezzulo, 2016), and health (e.g., Lim, Penrod, Ha, Bruce, & Bruce, 2018). Hand-initiated computer mouse movements are thus soundly used as motor traces of the mind (Freeman, Dale, & Farmer, 2011, p. 2). Using the mouse to uncover information also has a long tradition for studying how decisions are taken from a subset of information pieces, for instance in gambling tasks or economic decisions (Johnson, Payne, Bettman, & Schkade, 1989; Jasper & Shapiro, 2002; Franco-Watkins, Johnson, 2011).

While information must be accessed through mouse movements in the aforementioned studies, little information lies in the spatial organization of information across areas of interest in these types of experiments. On the contrary, spatial organization of information is a hallmark of construction rules in APMs. In this sense, our setup is closer to the kind of coupling between visual processes and mouse movements exploited in simple assistive technologies, such as the screen magnifiers embedded in most operating systems (allowing the user to attend to and move around a zoomed selection of the screen).

Research overview

In the present research, APMs were embedded in a graphical user interface, where the use of computer mouse movement allowed interactions to serve item exploration, rule inference, and rule testing. Item presentation was modified so that all useful information were not directly available to participants: only the matrix or the response alternatives was visible, never the two parts at the same time (contrary to classical APM displays). Participants were

required to switch from one part to the other by moving the computer mouse to the corresponding part of the item. In this sense, our design made it possible to exactly mimic the measures used to infer strategic behavior based on eye-tracking data: determining when participants are looking at the matrix versus the response alternatives, for what duration, and when they begin switching between the two. This procedure shares similarities with the design used by Mitchum and Kelley (2010), who had participants click to display the response alternatives (although the responses then remained on-screen and participants did not have to alternate between the matrix and possible responses).

Inferring the rules of the matrix via constructive matching should be achieved by spending time on the matrix and, once the missing pattern has been mentally constructed, switching to the response alternatives to select the matching answer (Bethell-Fox et al., 1984; Snow, 1978, 1980). Conversely, using response elimination necessarily requires switching between matrix and response alternatives to support the comparison process. Thus, the hypotheses that can be derived from such a design are identical to those of Vigneau et al. (2006) regarding strategy use: a higher proportion of time spent on the matrix part at the onset of item presentation and in total should reflect a constructive matching strategy, less time spent on the matrix part and more time on the response alternatives part should reflect response elimination. A lower number of switches between the two parts should be indicative of constructive matching, whereas a higher number of switches should indicate response elimination.

Contrary to eye-tracking studies, one of the core advantages of the present methodology is that these measures will necessarily reflect voluntary perceptual actions on the part of the participant during problem solving, and there will be no possible access to information in the hidden part of the item. Relatedly, another advantage of this approach is the possibility of manipulating where the participant is able to look at a given time. For example,

experimentally constraining time spent on the matrix at the onset of item presentation – by introducing a fixed delay before a voluntary switch can be performed to explore response alternatives – should favor adoption of the constructive matching strategy, at least for some individuals. This approach would take advantage of the flexibility of the graphical user interface, while complementing other methods devised to reduce response elimination and/or favor constructive matching (e.g., Arendasy & Sommer, 2013; Mitchum & Kelley, 2010).

The current research comprised three studies. First, a pilot study, conducted on a small sample, was designed to confirm that the proposed modification of the APM did not substantially change its psychometric properties. Experiment 1 verified that the indices of strategy use extracted based on mouse interactions were predictive of performance in the APM, in line with the study of Vigneau and colleagues (2006). Experiment 2 replicated the relation between strategy measures and performance, and also investigated how participant exploration of the item related to the solving process, by testing whether strategy use could be manipulated. We expected that constraining the time spent on the matrix at the onset of item presentation (as compared to an unconstrained condition identical to the interface version of the pilot study and Experiment 1) would improve APM performance by favoring the adoption of the constructive matching strategy.

Pilot study

This pilot study aimed at examining whether presenting the APM in their switch version, as compared to the Original version, altered the psychometric properties of the test[1]. In particular, we were interested in ensuring that the modified version elicited similar total score and completion time, had similar reliability and did not create differential item functioning.

[1] This study was part of a larger pilot study which included two additional versions of the interface. Given the exploratory nature of this pilot study, the between-participant nature of the version factor, as well as the fact that these other interfaces were not relevant in the context of this research, information and results are only reported for the two interface versions of interest.

Method

Participants and design. A sample of 36 undergraduate students (including 22 psychology students, 68% female, mean age = 20.9 years, SD = 2.88 years; and 14 computer science/mathematics students, 57% female, mean age = 20.2 years, SD = 0.98 years) participated in the study. Participants were randomly assigned to the Original or Switch condition of the APM, in exchange for course credit or a monetary reward (15 euros). Participants were from different university departments in an effort to recruit a more heterogeneous sample; the effect of field of study was included as a covariate in statistical analyses.

Apparatus and measures. Subjects completed the twelve-item short form of the APM, which has been shown to have psychometric properties similar to the long form (Arthur & Day, 1994). The Original condition of the APM was a computerized version of the task (identical to prior studies, such as Vigneau et al., 2006). The 3x3 matrix part was presented on the top half of the screen, and the eight alternatives were presented in two rows on the bottom part. Hence, all information was always visually available on the screen.

The Switch condition displayed exactly the same information in the same place as in the Original condition, but instead of making all information visually available on the screen, the matrix and response alternatives were not visible at the same time. Participants could display either part of the item by using the left mouse button to click in the top half or the bottom half of the screen. In other words, participants could alternate between the two parts by moving the mouse down and up the screen and by using the mouse click to display the part of interest, the other part being automatically hidden upon clicking (see Figure 1 for an example). For all items in the Switch condition, only the border of the missing pattern was displayed upon item presentation and kept visible at all times, a state that did not change until the participant actively took action with the computer mouse to display either the matrix or

response alternatives. There was no time limit in either condition and, in the Switch condition, participants could alternate as often as desired between the two parts. In both conditions, participants gave their response by clicking with the mouse on one of the eight response alternatives and by confirming their response with a keystroke, which allowed moving on to the next item.

All mouse interactions with the software were recorded. Records varied in frequency, depending on the speed and amplitude of mouse movements (with a maximum frequency of 200 Hz), with pixel level accuracy for the mouse position. Right and left mouse button clicks were also recorded. All these events were precisely associated with the part of the item hovered over and clicked on. Collected data thus provide accurate information regarding time spent on each part, number of alternations, selected response alternative, and time spent on each item. Total performance was automatically computed by the software upon completion. The Java source code and compiled software developed for this methodology can be found on the Open Science Framework (OSF; https://osf.io/um3wf/).

Procedure. After signing a consent form, all participants first performed two training items (items 10 and 6 from APM Set 1) with the Original version of the task. Participants were then randomly assigned to either the Original condition (n = 18) or the Switch condition (n = 18), and both groups performed two additional training items (items 8 and 9 from APM Set 1) with the task version corresponding to their experimental condition. They then completed the short form of the APM in their assigned condition.

Results and Discussion

To confirm that the switch version of the APM did not systematically hinder – or help – participants in performing the task, we first compared performance in the two conditions, controlling for students' major as a covariate. Total scores were highly similar between the Original condition (M = 8.55 out of 12, SD = 2.71) and in the Switch condition (M = 8.33,

SD = 2.74), with a nonsignificant difference between the two, $F(1, 32) = 0.08$, $p = .783$, $\eta^2_p = .00$. A bayesian analysis confirmed that there was moderate evidence in favor of the null hypothesis, $BF_{01} = 3.04$. Likewise, total time on task was similar between the Original condition (*M* = 661 seconds, *SD* = 354) and the Switch condition (*M* = 687 seconds, *SD* = 229), $F(1, 32) = 0.05$, $p = .828$, $\eta^2_p = .00$, $BF_{01} = 3.03$. Student's major, or its interaction with condition, also had no significant effect (all *p*s > .10.)

A related question was whether the switch version of the APM would exhibit differential item functioning: in other words, whether each item would demonstrate the same difficulty as in the Original version. Differential item functioning was tested with package *difR* (Magis, Béland, Tuerlinckx, & De Boeck, 2010) for *R* (R Core Team, 2017), using a Rasch model with Benjamini-Hochberg correction. None of the items demonstrated significant differential item functioning across the two versions, all *p*s > .30.

Lastly, we investigated whether the reliability of the Switch version would be similar to the Original version. Internal consistency was tested with Cronbach's alpha, which was acceptable for the Switch condition ($\alpha = .76$) and highly similar to the Original condition ($\alpha = .73$). The difference between these two coefficients, as tested with package *cocron* (Diedenhofen & Musch, 2016), was not significant, $\chi^2(1) = 0.06$, $p = .814$.

In sum, the results of the pilot study indicated that interface version did not strongly alter APM performance, completion time, reliability, or item properties. In Experiment 1, we used the same switch version of the APM in a much larger and homogeneous sample with the purpose of examining its suitability for assessing response strategies in matrix reasoning. To this end, we adapted the eye-tracking indices implemented by Vigneau and colleagues (2006) to this version of the task.

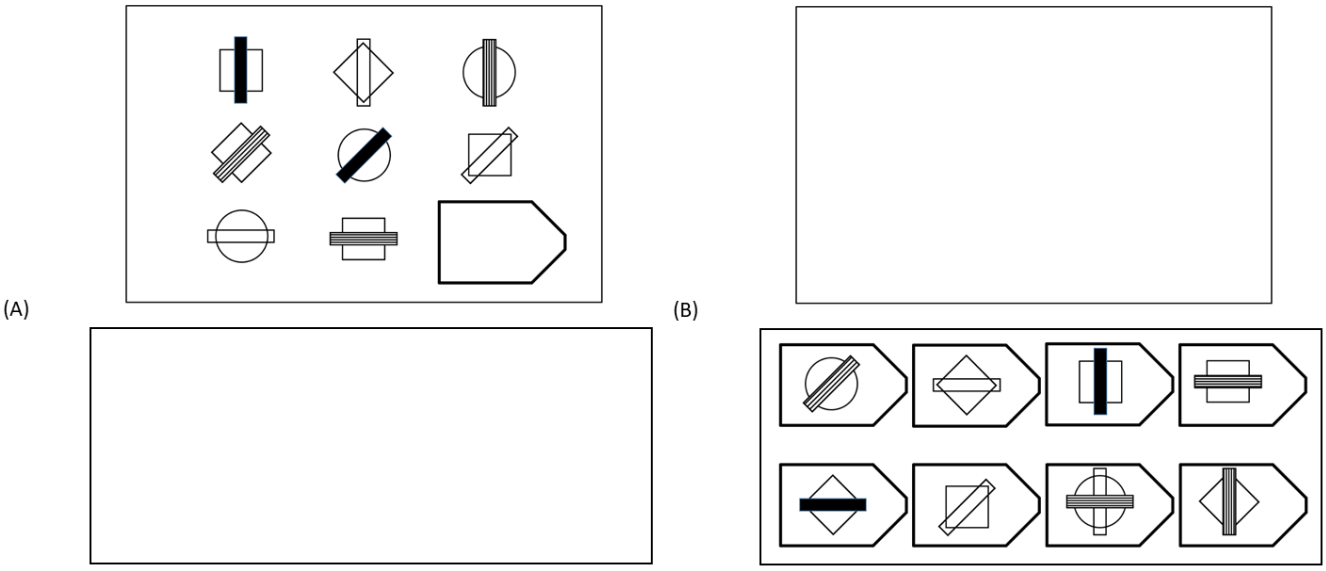

Figure 1. Illustration of the graphical user interface setup for a fictitious Advanced Raven Matrices (APM) item. This item uses two rules: constant in a row and distribution of three values. In the present research, either (A) only the matrix is visible (top part), or (B) only the response alternatives are visible (bottom part). Participants use the computer mouse to switch, when desired, from the top to the bottom part, and vice versa.

Experiment 1

Results of the pilot study indicated that the switch version of the APM retained similar psychometric properties as the Original version. Beyond this preliminary step, the main objective of the present study was to develop a methodology that could yield useful indices of strategy use: in other words, indices that would be predictive of performance in the APM, as is the case with measures used in eye-tracking (Vigneau et al., 2006). In Experiment 1, a larger sample of participants performed the Switch version of the APM. We computed the three strategy use measures of interest: proportion of time on the matrix vs. response alternatives, latency to first examination of response alternatives, and rate of alternation. The relation between these measures and total score was examined and compared to the results of Vigneau and colleagues (2006).

Method

Participants and design. A sample of 130 psychology students (87% female; mean age = 20.9 years; SD = 2.8 years) participated in exchange for course credit. Sample size was determined based on the lowest expected effect size for the whole experimental session. We doubled the sample size from Vigneau et al. (2006), and based on their results, we expected statistical power above 82% for the correlation between performance and time on alternatives, 88% for number of alternations, and above 99% for all other significant correlations in the original study. Even with an 80% CI around the estimated correlations (to take into account possible overestimation of the effects), a priori power remained above 83% for all these indicators of interest.

Apparatus and measures. All participants completed the Switch version of the APM, which was identical to the pilot study. This allowed the extraction of several mouse-tracking indicators: item latency (total time spent on the item), absolute time spent on the matrix, proportional time spent on the matrix (ratio of time spent on the matrix to total time spent on the item), absolute and proportional time spent on response alternatives, number and rate of alternations, and latency to first alternation (total time spent on the matrix before the first switch). All durations are expressed in minutes.

Procedure. The study was part of a larger experimental session (including achievement motivation items not reported here), with the whole procedure lasting approximately 45 minutes. The procedure was identical to the Switch condition of the pilot study, except that the APM were broken down into two sets of six items each, and participants completed two additional training items before the second set of six items (11 and 12 from APM Set 1 with the Switch version).

Analytic strategy. We ran two complementary series of analyses on the full data set comprising all indices. Firstly, Pearson correlation coefficients were computed between APM

performance and all indices of strategy use, for the purpose of direct comparison with the results of Vigneau and colleagues (2006). These correlations were computed on total performance aggregated by subject, and therefore neglected item-to-item, within-participant variance (which tends to be large in the APM, given the progressive nature of the items).

Secondly, as a way to take into account variability in the estimated parameters across both participants and items, mixed effect binomial models were fitted to the data to predict the success rate on each APM item for each participant. These analyses were performed using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015, version 1.1-19) in *R* software (R Core Team, 2017). Each index of strategy use was tested independently. The models for predicting success rate (R) included both an intercept and the effect of strategy index (S), as fixed effect parameters to be tested, as well as random parameters for both participant (P) and item (I), therefore using a formula of the form R~1+S+(1+S|P)+(1+S|I) using *R* notation.

Diagnostics for both series of analyses demonstrated no strong departure from models assumptions. Further details regarding analyses and diagnostic information are provided as Supplemental Material (available on OSF at https://osf.io/um3wf/). Analyzing the relation between the three key strategy indices and performance separately for each item yielded the same results as analyzing them at the aggregate level, as described above; details of item-level relations are also provided as supplemental material[2].

Results and Discussion

Our modified version of the APM retained acceptable reliability ($\alpha = .71$); total scores were normally distributed ($M = 6.16$, $ET = 2.67$, range = 1 – 12 out of 12, skewness = 0.06, kurtosis = -0.48). We began with a descriptive comparison of the measures collected in this experiment and in the study by Vigneau and colleagues (2006). The average success rate across all items was 51% (ranging from 89% for the first item down to 23% for the hardest

[2] Analyzing only the first set of six items yielded results similar to those reported in the paper, except that the weak negative correlation between number of alternations and performance appeared significant instead of nonsignificant, when using binomial mixed-effects model analysis only.

ones), when compared to 62% in the original study (ranging from 95% to 40%). Average time per item was also lower in our study (37 vs. 79 seconds). These differences do not seem critical, given that Vigneau and colleagues used a different (14-item) shortened version of the APM, and the composition of their sample was different.

As for strategy indices, we observed a drastically lower average number of alternations (2.3 vs. 21) and rate of alternation (0.08 vs. 0.29 / second) than Vigneau and colleagues. Despite these qualitative differences, the proportional time on matrix remained very similar (76% vs. 84%), as did (conversely) the proportional time on alternatives (24% vs. 16%). Lastly, the latency to first alternation was somewhat higher in our study (21 vs. 17 seconds). In other words, our design seems to have limited the number and rate of alternations between the matrix and response alternatives. These differences can be attributed both to the fact that we only recorded voluntary alternations between matrix and alternatives (which could be viewed as an advantage of this method); and to the fact that mouse interactions are more costly than eye movements (given that they involve increased muscular efforts and delays), thus discouraging alternations (which could be viewed as inconvenient). Given that participants spent the same amount of time on the matrix and response alternatives and did not wait longer before looking at response alternatives, it seems that the relative use of constructive matching and response elimination was not substantially changed, but the processes involved in implementing response elimination may have been different.

Reliability coefficients and results for the Pearson correlations between the various measures and total performance in the APM are reported in Table 1. All measures of interest were predictive of total accuracy in our study; correlations were generally comparable to or higher than those for the eye-tracking-based indices used by Vigneau and colleagues (2006). In particular, congruent with our expectations, performance was positively correlated with proportional time on matrix (r = .56) and negatively correlated with proportional time on the

response alternatives (r = -.56). This result is compatible with the hypothesis that these measures reflect the balance between constructive matching and response elimination[3]. Performance was also highly correlated with the other proposed index of constructive matching, latency to first alternation, r = .70. Conversely, a higher rate of alternation between matrix and response alternatives predicted lower performance, r = -.55, suggesting that this measure functioned well as an index of response elimination. The total number of alternations did not correlate with performance, but the alternation rate is more diagnostic as it controls for differences in response latency.

In sum, the analysis of bivariate correlations indicated that the three major indices of strategy use extracted from mouse interactions – proportional time on matrix vs. response alternatives, rate of alternation, and latency to first examination of the response alternatives – were all good predictors of performance in the APM (all $|rs| > .55$). An analysis of internal consistency also indicated that all three measures had excellent reliability in this sample (proportional time on matrix: $\alpha = .86$; latency to first examination of the responses: $\alpha = .86$; rate of alternation: $\alpha = .92$). The correlations between these three measures were also substantial (proportional time on matrix and latency to first examination of the responses: $r = .74$, $p < .001$; proportional time on matrix and rate of alternation: $r = -.79$, $p < .001$; latency to first examination of the responses and rate of alternation: $r = -.69$, $p < .001$). In other words, all three measures appear to be useful indices of strategy use. Given the high correlation and the conceptual relations between these three indices, a possible approach would be to compute a composite index of strategy use by averaging the three (after standardization); computing this composite index yielded a significant correlation with total accuracy but did not show much of an improvement over the three indices separately, r = .67,

[3] Note that these two indices are the inverse of each other in our study, given that proportional time on matrix and proportional time on response alternatives add up to 100%, which results in opposite correlations. This is not the case with eye-tracking studies, as participants sometimes look neither at the matrix nor at the response alternatives, but at an empty section of the display.

$p < .001$.

The final series of analyses used binomial mixed-effects models to confirm the relation between the various measures and performance, while controlling for participant and item random effects. The results are detailed in Table 1; for a more direct understanding of the effects, we also report the change in success rate for a mean value of the indices below. Confirming and extending the results of correlational analyses, these analyses indicated that the same three indices of strategy use were predictive of performance. Proportional time spent on the matrix and response alternatives had opposite significant effects, with a 5.8% increment in success rate for a 10% increase in time spent on matrix (b = +/-2.33, SE = 1.04, $\chi^2(1) = 4.24$, $p = .04$). Latency to first alternation was positively related to accuracy, with a 46% success rate increment per minute (b = 1.86, SE = 0.65, $\chi^2(1) = 8.12$, $p = .004$). Lastly, rate of alternations was negatively related to accuracy, with a 3.5% decrease for each additional alternation per minute (b = -0.14, SE = 0.04, $\chi^2(1) = 11.01$, $p < .001$).

Conversely, other measures not directly reflecting strategy use were not related to performance. The relation between item latency and accuracy was nonsignificant, with a 4.6% success rate increment per minute at the mean value of latency (b = 0.18, SE = 0.21, $\chi^2(1) = 0.72$, $p = .40$). Total time spent on the matrix had no significant effect on accuracy, with a 7.7% increment per minute on the matrix (b = 0.31, SE = 0.26, $\chi^2(1) = 1.41$, $p = .24$); the same was true for time spent on response alternatives, with a nonsignificant 7.5% decrement per minute, (b = -0.30, SE = 0.99, $\chi^2(1) = 0.09$, $p = .76$). Number of alternations also had no significant effect, with a decrease in 2% success rate per alternation (b = -0.08, SE = 0.07, $\chi^2(1) = 1.46$, $p = .23$). Again, this was the only difference from the results of Vigneau et al. (2006), though the number of alternations is less relevant than the alternation rate.

Table 1. *Relation between strategy indices and accuracy in the APM, for Vigneau et al. (2006) and for Experiment 1.*

	Vigneau et al. (N = 55)		Present Experiment 1 (N = 130)			
	Correlation		Correlation		Mixed model	
Predictors	*r*	*p*	*r*	*p*	*b*	*p*
Item latency	.03	.799	.63	<.001	0.18	.395
Time on matrix	.08	.537	.65	<.001	0.31	.236
Time on alternatives	-.25	.060	.29	<.001	-0.30	.763
Proportional time on matrix	.48	<.001	.56	<.001	2.33	.040
Proportional time on alternatives	-.44	<.001	-.56	<.001	-2.33	.040
Number of alternations	-.27	.050	.02	.835	-0.08	.226
Rate of alternations	-.43	.002	-.55	<.001	-0.14	<.001
Latency to first alternation	.41	.003	.70	<.001	1.86	.004

Note. *r* represents Pearson correlations between each index and total score in the task; *b* represents unstandardized slopes for the effect of each index on item-level accuracy, expressed as a log-odds ratio; *p* represents the corresponding *p*-values. Since the correlation estimates in Vigneau et al. (2006) are only known up to 2 decimals, only bounds on *p*-values can be provided here. The lower bound was kept for $p \geq .05$, and upper bound for $p < .05$.

To sum up, the results of Experiment 1 supported the relevance of the selected measures of strategy use - based on participant-initiated mouse movements to visualize either the matrix or response alterntives - as predictors of final APM performance. Going beyond correlational results, the findings in Experiment 1 were also stable over finer-grained analyses

taking participant and item-level variability into account, allowing for the results to be generalized to a larger population of APM items.

Critically, our indices of strategy use were generally comparable to indices collected in an eye-tracking study (Vigneau et al., 2006): subjects spent a similar amount of time on the matrix and response alternatives and waited for a similar duration before looking at the response alternatives. The only major discrepancy was that participants alternated much less frequently between the matrix and response alternatives with our modified version of the task. This result is not surprising, given that the motor cost of switching between the two parts of the item is certainly higher when the switch is performed using the hand and mouse than when it is performed using eye movements. We return to this point in the general discussion.

Likewise, the correlational pattern between our measures and performance was similar to the same correlations when computed from similar eye-tracking-based indices (Vigneau et al., 2006), and our method performed at least as well as eye-tracking. Aside from minor variations that could be attributed to limited sample size, the only major discrepancy between our results and those of Vigneau and colleagues is that item latency, time on matrix and time on alternatives were all positively correlated with performance in our study; but this is often the case with the APM (see Becker et al., 2016; Goldhammer et al., 2015; Perret & Dauvier, 2018), and it is more surprising that the same correlations were not significant in the study of Vigneau and colleagues. Given that these measures are not employed as indices of strategy use, and that indices of strategy use control for item latency, this is not a significant issue in this context.

Putting these findings into a broader perspective, one may consider that they demonstrate convergent validity with eye-tracking studies focusing on indices of strategy use. In other words, the use of interaction traces (be they eye-initiated as in eye-tracking or hand-mouse initiated as with our method) to explore problem-solving strategies in figural matrix

reasoning yields convergent conclusions, which suggests they represent a sound process-based method.

This consistent evidence supporting the relevance of the present interface for exploring strategy use in matrix reasoning notwithstanding, one major limitation of this experiment was the correlational nature of the design. To provide further evidence that user-interface interactions serve strategy use, Experiment 2 used an experimental design. We reasoned that if a higher proportion of time spent on the matrix (versus response alternatives) positively predicts performance because it favors a constructive matching strategy, then we could take advantage of the flexibility of our graphical user interface to manipulate the accessibility of response alternatives in order to influence performance. Since the time spent on matrix at the onset of item presentation is more directly reflective of a constructive matching strategy (hence the relevance of latency to first alternation), we could introduce a delay before participants could see the response alternatives part for the first time. This should influence strategy use and thus indirectly performance, at least for some participants and/or for some items (possibly in combination). Indeed, the delay should positively impact participants who spontaneously adopt but are not firmly set on the elimination strategy, and who would be induced to switch to constructive matching to quickly solve a given item.

On the contrary, participants who hold to the elimination strategy will simply be delayed in their resolution process. Therefore, in addition to the more common correlational analyses, analyses for Experiment 2 will also use mixed-effects models to appropriately take these inter-individual and inter-item differences into account, focusing on the interaction between the constraint imposed on participants and the time needed to solve the APM (item latency; see for instance Goldhammer, Naumann, & Greiff, 2015, signaling the importance of taking variations in item latencies across individuals and items into account in matrix reasoning). In Experiment 2, the results of correlational analyses between indices (as in

Experiment 1 and derived from Vigneau et al., 2006) and total APM score, performed for each condition separately, will be reported for consistency reasons and descriptive comparisons between Experiments 1 and 2.

Experiment 2

The main objective of Experiment 2 was to experimentally manipulate the strategies used by participants, by controlling the availability of the response alternatives through the computer interface. We contrasted two conditions, with either a constrained or unconstrained access to the response alternatives at the onset of item presentation (hereafter Constrained and Unconstrained conditions). Hiding response alternatives at the beginning of each item resolution imposes a longer latency to first examination of response alternatives. This should minimally delay – and possibly hinder – adoption of the response elimination strategy (for those who would like to look at the response alternatives before the imposed delay). Under these experimental conditions, and although participants cannot be forced to commit to a constructive matching strategy, trying to solve the item quickly requires processing of the matrix only. Participants willing to adopt the response elimination strategy need to wait for the imposed delay before engaging in the resolution process. We therefore expected differences in the effect of condition (constrained vs. unconstrained) depending on item latencies, themselves resulting from inter-individual differences in how participants react to the constraint.

Method

Participants and design. A sample of 64 psychology students (73% female; mean age = 20.7 years; SD = 2.1 years) participated in exchange for course credit and were randomly assigned to one of the two conditions. Sample size was determined based on an average effect size for the effect of interest (Cohen's *d* equivalent of 0.45) and a statistical power of 80%. Two participants in the constrained condition were removed from the sample

due to instructions misunderstanding (either selecting a response alternative before its content was made available, or not even selecting the matrix part before answering).

Apparatus and measures. One group of participants (n = 33) completed the switch version of the APM, which was identical to the switch condition of the previous studies. The other group of participants (n = 29) completed a similar switch version of the APM but were instructed after the four training items that response alternatives would only be made available after a fixed delay. Before this delay, participants could click on the response alternatives part, but would only see "locks" instead of possible answers (these attempted alternations were also recorded). Participants were not notified when the delay had elapsed, so that they could focus on the matrix resolution. The delay was adjusted for each item separately, since APM items vary widely in complexity: we used the third quartile of the latency to first alternation distributions from Experiment 1 (corresponding values for each item are reported as Supplemental Material, available on OSF at https://osf.io/um3wf/). The rationale behind this choice was to impact a large enough proportion of participants (75% of Experiment 1 sample), while avoiding an irritatingly long delay, in order to keep participants engaged in the task (i.e. shorter than average item latencies).

Procedure. The study was run as a stand-alone experimental session, with the whole procedure lasting approximately 30 minutes. The procedure was identical to the Switch condition of the pilot study.

Analytic strategy. To confirm earlier results and to check whether our experimental conditions indeed impacted the resolution of APM items, we first computed Pearson correlation coefficients between APM performance and all indices of strategy use, aggregated over items in each condition, for direct comparison with the results of Vigneau and colleagues (2006) and Experiment 1.

Focusing on success difference between conditions, a mixed-effects binomial model

was then fitted to the data for increased generalizability over items and participants, predicting the success rate on each APM item and participant. We expected strong inter-individual and inter-item differences in how the constrained interface would impact behavior and performance. Indeed, participants who spontaneously adopt the constructive matching strategy may be unaware of the change, if they switch to the response alternatives after the imposed delay. On the contrary, if they start looking for their constructed answer before the delay, they will have to wait for the locks to disappear (possibly going back to the matrix in the meantime). The correctness of such participants should thus be only marginally affected, but item latency may be increased.

At the other end of the strategy spectrum, participants who (feel the) need to adopt the response elimination strategy will have to wait for the full delay before even starting the resolution process. They will probably see their performance probably unchanged, but their item latencies greatly increased. Finally, participants who want to solve the items quickly yet are not set on a particular strategy will be forced to adopt constructive matching instead of response elimination at the beginning of the resolution process, due to our experimental manipulation.

Our model therefore required testing the effect of the interaction between item latency (T) and condition (C) on success rate (R). Using R formula notation, including random parameters for both participant (P) and item (I), the formula for the maximal model to be fitted is: R~C*T+(T|P)+(C*T|I)[4]. Estimating this model led to singularities, and the random structure of the model was reduced following Bates, Kliegl, Vasishth, & Baayen's (2015) recommendations to prevent incorrect parameter estimation. Since time is relative to the difficulty of the item, all item latencies were centered on the imposed delay (zero thus becoming the value at which response alternatives were made available for all items). This

[4] C*T being expanded to the intercept, individual effects of independent variables and their interaction (1+C+T+C:T)

allowed descriptive and inferential statistics to be more consistent, while making it possible to satisfy model assumptions (especially distributions of random effects). We also removed all trials where a response was given before the disappearance of the locks in the constrained condition, since they were necessarily random and could bias our estimates (three trials removed in addition to those of the two participants previously removed from the sample). The statistical results were robust to both changes, but their interpretation was made easier. Further details regarding analyses and diagnostic information are provided as Supplemental Material (available on OSF at https://osf.io/um3wf/).

Results and Discussion

Our modified version of the APM retained nearly acceptable reliability in the unconstrained condition ($\alpha = .69$), with total scores roughly normally distributed ($M = 7.48$, $ET = 2.60$, range = 2 – 12 out of 12, skewness = -0.07, kurtosis = -0.87). Reliability in the constrained condition was reduced ($\alpha = .46$), with a roughly symmetrical score distribution ($M = 7.93$, $ET = 2.00$, range = 4 – 11, skewness = -0.22, kurtosis = -1.14).

As a preliminary analysis, we first examined whether indices of strategy use in the Unconstrained condition demonstrated the same relations to performance as in Experiment 1. Results for the Pearson correlations between the various measures and total performance in the APM are reported in Table 2, computed separately for each condition. In the Unconstrained condition, we replicated the pattern found for most indices in Experiment 1. Again, item latency, time on matrix, proportional time on matrix and latency to first alternation were all significantly and positively correlated with total accuracy. Proportional time on alternatives and rate of alternations were again found to be significantly and negatively correlated with accuracy. Time on alternatives and number of alternations were weakly correlated with accuracy, and this time these correlations were both non-significant. The reduced significance compared to Experiment 1 was probably due to the lower sample

size in each condition.

In the Constrained condition, item latency, time on matrix, rate of alternations and latency to first alternations followed the same pattern as in the Unconstrained condition, but with lower values; all were nonsignificant. This logically reflects the influence of the initial period with no access to response alternatives, since not only are raw temporal measures shifted, but any activity oriented towards response alternatives during this period is also recorded and adds noise to the data. The only exceptions were time on alternatives and number of alternations, which became negatively - yet nonsignificantly - correlated with accuracy: this is expected, since alternations and time spent on response alternatives during the initial matrix-only period clearly become counterproductive in item resolution, and should be used merely to check the availability of response alternatives (given that only locks are displayed with the matrix also hidden). For the same reason, correlations between proportional times (on matrix and on alternatives) and accuracy are increased and remain significant. Shifting time measures by the imposed delay would not help here, since data were aggregated at the participant level in order to compute the correlations.

The main objective of this experiment was to test the difference in accuracy between conditions, in interaction with item latencies. Binomial mixed-effects model analysis indicated that the interaction between condition (Constrained vs. Unconstrained) and item latency (centered on imposed delay for each RAPM item) was significant (b = -0.013, SE = 0.006, $\chi^2(1) = 4.80$, p = .03). The condition effect at the time when responses alternatives appear (minimal item latency in Constrained condition) was also significant (b = 0.66, SE = 0.33, $\chi^2(1) = 3.96$, p = .047), with a higher success rate in the Constrained condition (M = .78, SE = 0.07, 95% CI [.62, .89]) than in the Unconstrained condition (M = .65, SE = 0.09, 95% CI [.47, .80]). The

Table 2. *Relation between strategy indices and accuracy in the APM, for Experiment 2.*

Condition	Unconstrained (n = 33)		Constrained (n = 29)	
Predictors	*r*	*p*	*r*	*p*
Item latency	.49	.004	.18	.365
Time on matrix	.51	.003	.28	.140
Time on alternatives	.23	.060	-.31	.097
Proportional time on matrix	.37	.001	.46	.012
Proportional time on alternatives	-.37	.001	-.46	.012
Number of alternations	.18	.050	-.06	.756
Rate of alternations	-.37	.035	-.13	.517
Latency to first alternation	.52	.002	.27	.158

This beneficial effect of the interface constraint on success rate when the response alternatives first appeared, as represented in Figure 2, reflects the higher proportion of participants adopting a constructive matching strategy - which is the only one of the two prototypical strategies that can lead to early responses in this condition. The beneficial effect of condition nevertheless gets reversed when participants validated their answer more than 49 seconds after the response alternatives appeared (16% of all trials; see Figure 2). This reversal should be the logical consequence of longer item latencies for participants failing to successfully solve the APM item with the constructive matching strategy before response alternatives appear, and then switching to the response elimination strategy. Given that response elimination is negatively correlated with success rate, the overall proportion of correct responses for high item latencies decreases, down to partly compensating the initial benefit for quick responders.

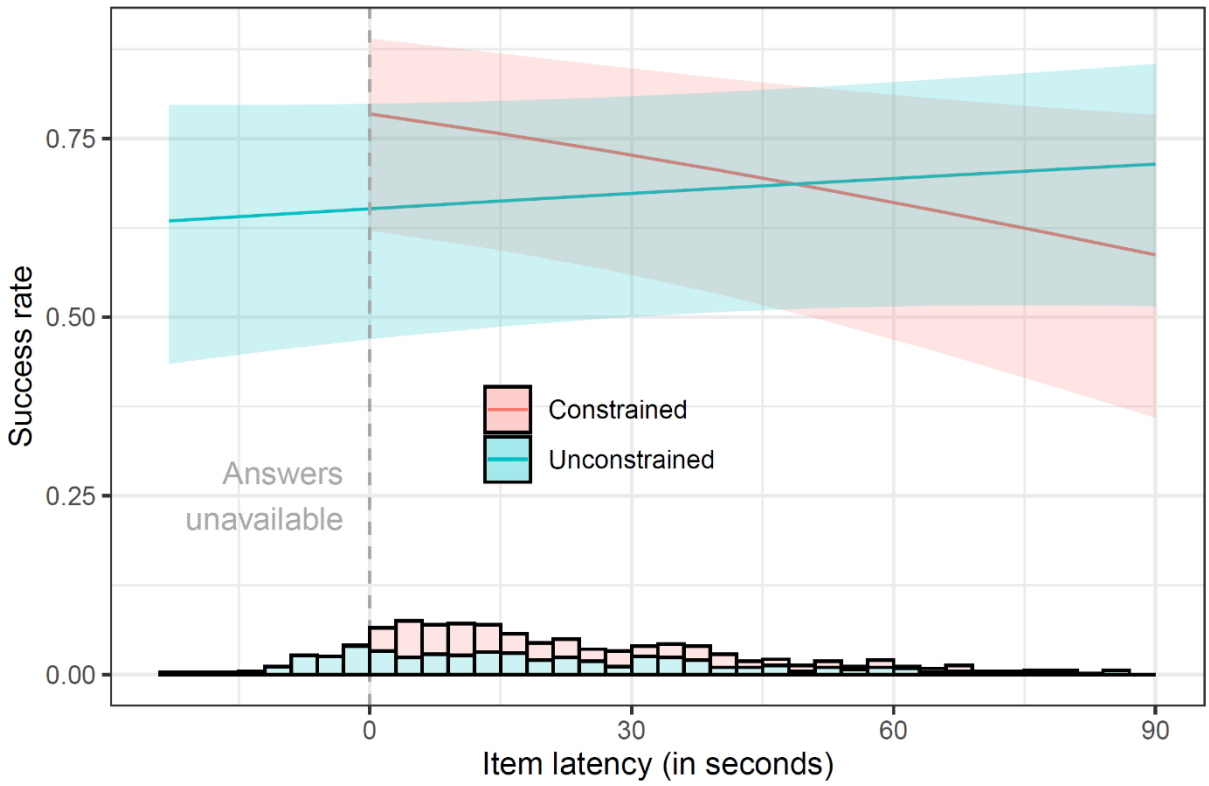

Figure 2. Success rate estimates from binomial mixed-effects model analysis in Experiment 2 as a function of condition and item latency. Item latencies were centered on imposed delay for each item, with no response possible before this delay in the Constrained condition. Histogram of item latencies overlaid at the bottom.

General discussion

The goal of the present research was to test the methodological and theoretical relevance of embedding APM in a newly designed graphical user interface, where hand-initiated computer mouse movements allowed user-interface interactions to serve item exploration, rule inference, and rule testing. Based on previous eye-movement studies and active vision principles, mouse movements were expected to provide insight into reasoning strategies.

The results for the pilot study indicated that this procedure (modifying the display of APM items, with only the matrix or the response alternatives being visually available at the same time) did not impair performance or completion time, did not elicit marked differences

in item functioning, and did not decrease the reliability of the task when compared to the original APM. Experiment 1 indicated that indices of strategy use based on participant-initiated switches between the matrix and the response alternatives were highly predictive of performance, like similar eye-tracking-based indices (Vigneau et al., 2006). These results were stable over finer-grained analyses taking participant and item-level variability into account. Experiment 2 manipulated the information made available at the beginning of each item (hiding response alternatives), only allowing for the use of constructive matching; the results confirmed that the necessarily larger proportion of participants adopting a constructive matching strategy for low item latencies led to an increase in performance, with a decrease and then reversal of this tendency for higher item latencies.

In short, the present results suggest that having participants use the mouse to switch between the matrix and response alternatives does not degrade the quality of the task and provides indices consistent with the use of the two major reasoning strategies – constructive matching and response elimination. In spite of differences in effectors (hand versus eye) and display (only one part of the item visible versus full display), the relations between indices of strategy use and APM performance were strikingly similar in Experiment 1, in the unconstrained condition in Experiment 2, and in the eye-tracking results of Vigneau et al. (2006). It is noteworthy that for the three key indices of strategy use (proportion of time on matrix, latency to first examination of the response alternatives, and rate of alternation), correlations were of a higher magnitude in the present experiment than in prior results using eye-tracking (Vigneau et al., 2006). This was especially the case for latency to first alternation, which was a significantly better predictor of performance in Experiment 1 data (Fisher *r*-to-*z* test: $p = .008$).

The present set of findings therefore provides evidence that deliberate item exploration supported by computer mouse movements represents a good methodology for accessing

relevant problem-solving strategies. When compared to eye-tracking, this methodology has the additional advantages of user-friendliness, implicitness, portability, and easy implementation of interaction constraints. For these reasons, this method would be particularly well suited to the investigation of reasoning strategies in populations for which the use of eye-tracking could be difficult, such as children, and in populations more readily reachable through online studies. Our findings also provide converging evidence that indices collected from eye-tracking do reflect part of the reasoning processes and strategies of the participants.

The one major discrepancy between our findings and eye-tracking was the drastically lower number of alternations, on average, between the matrix and response alternatives. This decrease is logical given the higher motor cost required to view response alternatives with the hand (which requires moving the mouse to the other part of the screen) than with an eye saccade. It is unclear to what extent this decrease is problematic. On one hand, it could be viewed as a benefit of our method: the results reflect only deliberate exploration of items on the part of participants, potentially increasing the specificity of the measure. The alternation rate correlated with the other two strategy indices, and had a large negative correlation with performance in Experiment 1 - at least as high, in fact, as the one obtained with eye-tracking - suggesting that this measure still reflected response elimination at least as well.

A possible issue would be if the lower alternation rate indicated that our design discouraged response elimination, but this did not appear to be the case. The fact that our modification did not increase average performance, that participants spent on average the same amount of time on the matrix and response alternatives, and that they waited for the same duration before looking at response alternatives, all converge to suggest that our design did not lead them to use *less* response elimination. Instead, the lower number of alternations suggests that participants used response elimination *differently*. It could be the case, for

example, that they considered all response alternatives less systematically, or that they made a greater effort to remember features of the matrix before looking at response alternatives (or vice versa) to minimize the number of required alternations. In short, our method appears to be appropriate to assess the relative contributions of constructive matching and response elimination, but it also indicates that participants may implement response elimination in different ways.

This effect of the design on alternation rate raises at least two questions for future research. The first question is the interpretation of the three key indices of strategy use. Proportion of time on the matrix, latency to first alternation, and alternation rate are usually considered as three (interchangeable) measures of the relative use of constructive matching and response elimination; but the fact that our design drastically reduced one of the three indices without affecting the inter-correlations of the three indices, or their relation with performance, shows that they may in fact tap into different aspects of strategic behaviors.

The other question is to what extent our modified design changes the nature of the task, in terms of the cognitive processes required for successful performance. The role of working memory, in particular, may be different if participants have to remember features of the no-longer-presented-matrix when looking at response alternatives. Working memory capacity is already strongly correlated with Raven's matrices (Ackerman et al., 2005), so we believe that enhancing its role in the task is not necessarily a serious issue. However, our design may blur the relation between working memory capacity and strategy use. For example, a high working memory capacity is usually associated with more constructive matching and less response elimination (Gonthier & Thomassin, 2015; Gonthier & Roulin, 2019), but with our design it could also be the case that participants with a high working memory capacity need fewer alternations to use response elimination. This approach should thus be used with caution to test the relation with working memory.

A possible avenue for future research pertains to the possibility of modifying the interface even further. Experiment 2 focused on the indirect manipulation of latency to first alternation (since we could not directly manipulate the strategies adopted by participants), but other indices could be manipulated, for instance by putting a limit on the number or rate of alternations for each item, thereby limiting the adoption of response elimination while not forbidding access to response alternatives at any point in time. Interface modifications could also be used for a finer-grained investigation of response processes, beyond the use of the two classical strategies. There are at least two ways to do this.

Firstly, rule inference in the APM predominantly relies on uncovering regularities in rows through pairwise comparison of cells in the same row (Carpenter et al., 1990; Hayes et al., 2011). To investigate this process, information displayed in the matrix part may be restricted to visual accessibility of a single row. That is, using computer mouse movements, participants could visually display only one row after another in the matrix part. Secondly, successful rule inference seems to depend on the consistency of this scan pattern across successive cells of the matrix: performance is higher in subjects who distribute their attention more evenly across all cells (Vigneau et al., 2006), and in subjects who systematically follow an ordered sequence of fixations across adjacent cells and rows (Hayes et al., 2011). This could be investigated using a different modified version of the interface allowing users to select and view any of the cells of the matrix with the computer mouse. With this alternative interface, it would be possible to unveil uncommon exploration and inference strategies, with undesired peripheral visual information in the adjacent cells remaining hidden unless the participant clicks on them.

Such alternative interfaces have the potential to provide an in-depth understanding of how item display manipulation may facilitate or impair matrix reasoning, and to uncover uncommon exploration and inference strategies, as information that is usually available in

classical APM displays is hidden. Implementing such interfaces is not quite straightforward: they would be more invasive than the solution presented in the current study, so that they could impact performance to a greater extent. On one hand, they may harmfully constrain information accessibility, hence impairing rule inference and testing in matrix reasoning - for example by placing such high demands on working memory for features of the no-longer-displayed parts of an item that complex items become very difficult to solve. On the other hand, extrapolating from the results of Experiment 2, increasing the saliency of row-dependent regularities by constraining access to a single row at a time may act as a scaffolding technique, hence providing support and guidance in the problem-solving process, and facilitating rule inference.

In practice, we did perform pilot testing for both possibilities outlined above. Participants reported that these other design changes interfered with the resolution of the task, so these versions were not further explored and the corresponding results are not reported here, but we believe these alternative solutions deserve to be explored in greater depth. In our opinion, further developing these types of manipulations of item displays in matrix reasoning tasks represents an important avenue for future research, one which has the potential to overcome limitations inherent in correlational studies focusing on problem-solving strategies in matrix reasoning. Several studies have fruitfully used these types of manipulations in the past (e.g., Arendasy & Sommer, 2013; Mitchum & Kelley, 2010; Duncan, Chylinski, Mitchell, & Bhandari, 2017; Rozencwajg & Corroyer, 2001).

One last possible application concerns interpretation of intelligence scores for a given individual. Of particular interest, the present setup provides easy access to mouse movements signaling the use of the more effective constructive matching strategy (at least in terms of final performance), or the use of the less effective response elimination strategy. Researchers and practitioners interested in using the APM as a diagnostic tool may further wish to use

mouse data to examine, for a given individual, the predominance of one or the other of these strategies. The excellent internal consistency coefficients observed for the three measures of strategy use ($\alpha > .85$) indicate that this could be a reliable approach to assessing determinants of reasoning performance.

Beyond the APM, the advantages of computer mouse data – previously favored in experimental psychology for forced-choice categorization tasks – in terms of user-friendliness, implicitness, portability, and controllability may be expanded to more diverse paradigms and setups, while also offering methodological perspectives to manipulate exploration and inference processes in matrix reasoning.

Open Practices Statement

The Java application for the interfaces (working demo included), data, and analysis scripts for all studies are available on OSF (https://osf.io/um3wf/).

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30-60. doi: 10.1037/0033-2909.131.1.30

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function, 218*(3), 611–643. doi: 10.1007/s00429-012-0475-5

Arendasy, M. E., & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence, 41*(4), 234-243. doi: doi.org/10.1016/j.intell.2013.03.006

Arthur, W., & Day, D. V. (1994). Development of a Short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54*(2), 394–403. https://doi.org/10.1177/0013164494054002013

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. *Nature, 221*(5184), 963.

Clutch (2017). Clutch 2017 Consumer App Onboarding UX Survey. Retrieved from https://clutch.co/app-developers/resources/mobile-app-onboarding-survey-2017

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1)*, 1-48. doi:10.18637/jss.v067.i01.

Bernstein, N. A. (1967). The co-ordination and regulation of movements: Conclusions towards the Study of Motor Co-ordination. *Biodynamics of Locomotion*, 104-113.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238. https://doi.org/10.1016/0160-2896(84)90009-6

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on raven's advanced progressive matrices. *Journal of Applied Psychology, 91(4)*, 979. http://dx.doi.org/10.1037/0021-9010.91.4.979

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Crossley, S., Duran, N. D., Kim, Y., Lester, T., & Clark, S. (2018). The action dynamics of native and non-native speakers of English in processing active and passive sentences. *Linguistic Approaches to Bilingualism*. https://doi.org/10.1075/lab.17028.cro

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*, 51–60.

Duncan, J., Chylinski, D., Mitchell, D. J., & Bhandari, A. (2017). Complexity and compositionality in fluid intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(20), 5295–5299. doi:10.1073/pnas.1621147114

Franco-Watkins, A. M., & Johnson, J. G. (2011). Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. *Judgment & Decision Making*, 6(8).

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241. https://doi.org/10.3758/BRM.42.1.226

Freeman, J. B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2*, 59. https://doi.org/10.3389/fpsyg.2011.00059

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*(4), 502–517. https://doi.org/10.1177/0956797615627418

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in cognitive sciences*, 16(2), 129-135.

Gibaldi, A., Vanegas, M., Bex, P. J., & Maiello, G. (2017). Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods*, *49*(3), 923–946. doi: 10.3758/s13428-016-0762-9

Goffart, L., Bourrelly, C., & Quinton, J.-C. (2018). Neurophysiology of visually-guided eye movements: Critical review and alternative viewpoint. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.00402.2018

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*(1), 21-40. doi:10.3390/jintelligence3010021

Gonthier, C., & Roulin, J.-L. (2019). Intra-individual strategy shifts in Raven's matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*, *149*(3), 564-579. doi:10.1037/xge0000660

Gonthier, C., & Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on Raven's matrices. *Journal of Experimental Psychology: General*, *144*(5), 916–924. https://doi.org/10.1037/xge0000101

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), 10-10.

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence*, *40*(5), 427–438. https://doi.org/10.1016/j.intell.2012.06.001

Jasper, J. D., & Shapiro, J. (2002). MouseTrace: A better mousetrap for catching decision processes. Behavior Research Methods, *Instruments, & Computers*, 34(3), 364-374.

Johnson, E. J., Payne, J. W., Bettman, J. R., & Schkade, D. A. (1989). Monitoring information processing and decisions: The Mouselab system (No. TR-89-4-ONR). Duke Univ Durham NC Center for Decision Studies.

Kelso, J. S., Fuchs, A., Lancaster, R., Holroyd, T., Cheyne, D., & Weinberg, H. (1998). Dynamic cortical activity in the human brain reveals motor equivalence. *Nature*, *392*(6678), 814.

Kietzmann, T. C., Geuter, S., & König, P. (2011). Overt Visual Attention as a Causal Factor of Perceptual Awareness. *PLoS ONE*, *6*(7), e22614. https://doi.org/10.1371/journal.pone.0022614

Kirsh, D., & Maglio, P. (1994). On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, *18*(4), 513–549. https://doi.org/10.1207/s15516709cog1804_1

Lim, S.-L., Penrod, M. T., Ha, O.-R., Bruce, J. M., & Bruce, A. S. (2018). Calorie labeling promotes dietary self-control by shifting the temporal dynamics of health-and taste-

attribute integration in overweight individuals. *Psychological Science*, *29*(3), 447–462. https://doi.org/10.1177/0956797617737871

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42(3)*, 847-862. doi:10.3758/BRM.42.3.847

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences*, 102(29), 9995-9996.

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition*, *15*(3), 620–627. https://doi.org/10.1016/j.concog.2006.01.001

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(3)*, 699. https://doi.org/ 10.1037/a0019182

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134). https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

O'Regan, J. K. (1992). Solving the" real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *46*(3), 461.

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(05), 939–973. https://doi.org/10.1017/S0140525X01000115

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of ‘mudsplashes'. *Nature*, *398*(6722), 34.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170–4175. https://doi.org/10.1073/pnas.1415250112

Philipona, D., O'regan, J., Nadal, J. P., & Coenen, O. (2004). Perception of the structure of the physical world using unknown multimodal sensors and effectors. In *Advances in neural information processing systems* (pp. 945-952).

R Core Team. (2017). R: A Language and Environment for Statistical Computing (Version 3.4.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raven, J. C., & Court, J. H. (1998a). Manual for Raven's progressive matrices and vocabulary scales. Section 4: The advanced progressive matrices. *San Antonio, TX: Psychological Corporation*.

Raven, J. C., & Court, J. H. (1998b). *Raven's progressive matrices and vocabulary scales*. Oxford pyschologists Press.

Régner, I., Smeding, A., Gimmig, D., Thinus-Blanc, C., Monteil, J. M., & Huguet, P. (2010). Individual differences in working memory moderate stereotype-threat effects. *Psychological Science, 21(11),* 1646-1648. https://doi.org/10.1177/0956797610386619

Rozencwajg, P., & Corroyer, D. (2002). Strategy development in a block design task. *Intelligence*, *30*(1), 1–25. doi:10.1016/S0160-2896(01)00063-0

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Smeding, A., Quinton, J.-C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. *Journal of Personality and Social Psychology*, *111*(6), 817–834. https://doi.org/10.1037/pspa0000063

Snow, R. E. (1980). Aptitude processes. *Aptitude, Learning, and Instruction*, *1*, 27–63.

Snow, R. E, (1978). Eye fixation and strategy analyses of individual differences in cognitive aptitudes. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 299-308). New York: Plenum Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). From The Cover: Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. https://doi.org/10.1073/pnas.0503903102

Spivey, Michael J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, *36*(6), 702–710. https://doi.org/10.1016/j.intell.2008.04.004

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261–272. https://doi.org/10.1016/j.intell.2005.11.003

Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

Wechsler, D. (2008). *WAIS-IV administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Wiley, J., & Jarosz, A. F. (2012). How Working Memory Capacity Affects Problem Solving. In *Psychology of Learning and Motivation* (Vol. 56, pp. 185–227). Elsevier. https://doi.org/10.1016/B978-0-12-394393-4.00006-6

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, *50*(1), 160–181. https://doi.org/10.3758/s13428-018-1127-3