




Metacognitive experience on Raven's matrices versus insight problems

Adam Chuderski¹  · Jan Jastrzębski¹ · Bartłomiej Krocze¹ · Hanna Kucwaj¹ · Michał Ociepka¹

Received: 17 December 2019 / Accepted: 17 July 2020/Published online: 27 July 2020
© The Author(s) 2020

Abstract

Participants rated Intuition, Suddenness, Pleasure, and Certainty accompanying their solutions to items of a popular fluid intelligence test – Raven's Advanced Progressive Matrices (RAPM) – that varied from easy (around 80% correct) to difficult (around 20% correct). The same ratings were collected from four insight problems interleaved with RAPM. Suddenness and Certainty substantially decreased from easy to difficult matrices (Pleasure strongly overlapped with Certainty). In easy matrices, subjective experience matched that observed during insight problems, suggesting the highly fluent processing resulting in vivid and univocal solutions. By contrast, processing difficult matrices seemed to involve effortful incremental combination of complex information that yielded uncertain outcomes, resembling full-blown analytic problems. Only Intuition, generally rated low, was unaffected by RAPM difficulty. These results suggest that RAPM constitutes a heterogeneous test, with easy vs. difficult items involving relatively distinct types of processing. This novel knowledge can help in understanding the processes underlying solving Raven's matrices. The study also contributes to the understanding of the validity of subjective ratings as measures of metacognition.

Keywords Raven's Matrices · Fluid intelligence · Metacognitive ratings · Insight

Introduction

Fluid intelligence – the ability to solve novel problems with abstract reasoning – is a proxy for general cognitive ability (McGrew, 2009) that predicts multiple socioeconomic variables, including professional success, income, and health (Deary, 2012). Fluid intelligence strongly

The order of authors is alphabetical

✉ Adam Chuderski
adam.chuderski@uj.edu.pl

¹ Institute of Philosophy, Jagiellonian University, 52 Grodzka St, 31-044 Kraków, Poland

links with cognitive capacities, such as attention and memory (Shipstead et al., 2014; Unsworth et al., 2014), and predicts learning potential, skill acquisition, and knowledge use (Colom & Flores-Mendoza, 2007; Deary et al., 2007; Demetriou et al., 2019; Hattie, 2009; Primi et al., 2010; Ohtani & Hisasaka, 2018; Roth et al., 2015).

A popular method for assessing fluid intelligence in students and adults consists of geometric matrices, such as Raven's Advanced Progressive Matrices (RAPM; Raven, 1938; Raven et al., 1998). RAPM includes 36 test items. Each item comprises a matrix figure of 3 rows and 3 columns with the bottom right-hand cell missing. The remaining eight cells contain geometric shapes. The features of these shapes may change across rows and columns according to several rules (see Carpenter et al., 1990). One relatively simple rule consists of a quantitative increment of a feature across either rows or columns, such as size, grid, rotation, the number of dots or stripes (see left panel of Fig. 1 for the box filling that gets darker from left to right, and the box orientation that changes from top to bottom). Another rule involves the distribution of two or three values across rows and columns (see the small, medium, and large rectangles distributed in the left panel of Fig. 1). Relatively difficult rules implement various variants of juxtaposition, superimposition, or subtraction for two figures or their elements, resulting in the third one (see right panel of Fig. 1 for the addition of grey lines and the intersection of black lines). The task is to identify all the rules governing shapes in the eight cells, and to infer the missing ninth cell. For each matrix, participants are required to choose the one and only correct option out of eight alternatives. The seven remaining options differ in how close they are to the correct option (Raven et al., 1998).

Despite 80 years of application of RAPM, processes involved in solving RAPM are still elusive. The aim of the present study was to broaden our understanding of performance in RAPM by collecting and analyzing the metacognitive and affective reports that reflect subjective experience accompanying the problem-solving process in RAPM items at varying levels of difficulty.

Cognitive processing on RAPM

A number of methods have been applied to understand the cognitive mechanisms that can be presumably involved in solving the RAPM items. One method consists of looking for

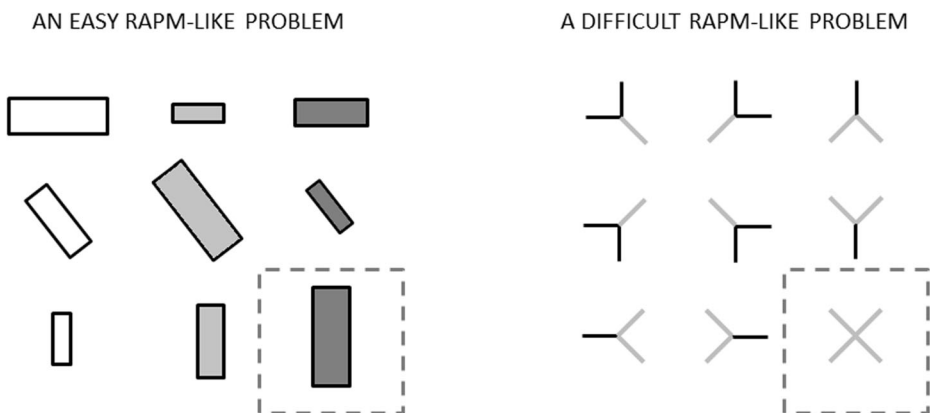


Fig. 1 Two exemplary RAPM-like items: an easy (left) and a difficult one (right); the correct response (here in a dashed rim) is missing and has to be selected out of eight options

elementary cognitive variables (e.g., various perceptual, attentional, memory, and learning capabilities of individuals) that correlate with scores on RAPM (typically, the scores are the number of items solved correctly). This line of research has indicated that the strongest predictor of individual RAPM scores (Oberauer et al., 2005) is working memory (WM) capacity. WM capacity (WMC) is most often defined as the number of objects a participant can actively maintain and transform in the mind, especially under concurrent processing and distraction (see Cowan, 2016). Other ways to operationalize WMC include the number and/or precision of objects' features that can be encoded, and the complexity of relations that bind the objects and features together (see Skrzypulec & Chuderski, 2020).

Another method to explore the cognitive mechanisms underlying performance in RAPM is to study the effects of varying the contents of a RAPM item. For instance, Primi (2001) manipulated the perceptual complexity of figures in RAPM, the number of these figures, and the number of rules that governed them. He found that perceptual complexity most strongly affected solution accuracy, followed by the number of figures and the number of rules. Meo et al., (2007) reported that matrices with salient elements were easier to solve, also suggesting that the proper decoding of rules from perceptual organization of a matrix may be crucial.

Eye tracking helped to identify strategies that people apply to RAPM (Bethel-Fox et al., 1984; Carpenter et al., 1990; Hayes et al., 2011; Jarosz & Wiley, 2012; Vigneau et al., 2006). This approach revealed that some participants systematically scan the matrix, mentally construct a likely solution, and only then compare it with the response bank (so-called constructive matching). Other people process the matrix less systematically, frequently switch to the response bank, and very early test which options can be rejected (so-called response elimination). Constructive matching yields higher scores than response elimination. The latter strategy is more frequently used by people low in WMC, and as a result their attention is more easily captured by distracting options present in the response bank (Jarosz & Wiley, 2012). The constructive strategy can be induced by making erroneous response options highly similar to the correct option, because no potential option can be easily eliminated as obviously wrong (Arendasy & Sommer, 2005). The eliminative strategy can also be discouraged by removing the response bank and requiring the participant to construct a response (Becker et al., 2016), as well as by dedicated training (Hayes et al., 2015) and instruction (Loesche et al., 2015).

Finally, strategy use has also been probed by means of Likert-like subjective ratings (Gonthier & Tommasin, 2015; Jastrzębski et al., 2018; Mitchum & Kelley, 2010). In this method, several statements describe ways of behavior typical for a given strategy, and participants have to report the extent to which they behaved in each way (e.g., "I switched between the problem and the response options: frequently – at times – rarely – never."). These studies more or less supported eye-tracking results. Less constrained variants of self-report comprise verbal protocols, that is, verbalizing thoughts during solving a test. For instance, Jarosz et al., (2019), by properly structuring the resulting protocols, have identified another strategy used in RAPM (so-called isolate-and-eliminate) that merged aspects of (simple) constructive matching and of response elimination.

However, all the studies employing subjective reports focused on strategy use. In order to better understand RAPM processing, this study aimed at a more comprehensive use of subjective ratings beyond sheer strategies (i.e., beyond the stage of searching the solution), and also examined subjective experience that occurred just after a solution had been found.

Subjective experience in insight problems

Subjective ratings pertaining to the solution found have recently been used in insight problem solving research. Insight problems are special types of problems designed to induce creative yet convergent solutions (see Batchelder & Alexander, 2012; Chu & MacGregor, 2011; Kounios & Beeman, 2014). Their instructions suggest a typical problem representation and solution strategy; however, pursuing such a line of reasoning actually results in impasse (i.e. being stuck in the solving process with no idea of how to move on). For instance, one needs to transform an incorrect equation in Roman numerals made of matchsticks – such as ‘VI = VI + VI’ – into a correct equation by moving just one matchstick (with no adding or removing of matchsticks allowed). After certain time, one comes to realize that no sum or difference on the right side will give a valid number on the left side (different moves just lead to a sequence of incorrect equations, like ‘VII = VI + V’, etc.) What one must recognize in these matchstick algebra problems is that equations do not necessarily include only one equation sign, but can include two, resulting in the tautology ‘VI = VI = VI’ (Knöblich et al., 1999). Also, having been instructed to use exactly six pencils (without breaking or bending any) to construct four triangles (a classic spatial problem), one might spend hours searching for a solution in the 2D plane until coming up with a tetrahedron (Katona, 1940). And, given 1 min to solve the following classic math problem “A man went to zoo and saw the giraffes and ostriches, which altogether had 30 eyes and 44 legs. How many animals were there?”, one can run out of time trying to calculate the giraffes and ostriches separately, while the rapid correct answer is 15 (the question is about the total number of animals, and each animal has exactly two eyes).

Although a number of studies analyzed unstructured verbal protocols during insight problem solving (e.g. Fleck & Weisberg, 2004, 2013; Schooler et al., 1993), of particular interest here are more structured reports in the form of questionnaires. A pioneer of this method, (Metcalf 1986; Metcalfe & Weibe, 1987) examined a unidimensional index of subjective experience: the feeling of warmth indicating the perceived distance to a solution. The index yielded a quite different time course in insight problems vs. analytic problems, the latter involving well-defined steps (e.g. algebra problems). While analytic problems yielded a steady, gradual increase in the subjective feeling of warmth, in insight problems, there was no increase until the very last moment, when the solution was actually found. That the solutions popped out suddenly and unexpectedly was interpreted as suggesting the special nature of insight problems, as compared to other types of problems (for a critique, see Weisberg, 1992).

Recently, probing subjective experience has gained popularity, yielding a considerable outburst of important findings on the relation between subjective ratings, problem types and conditions, and the actual problem-solving performance. Whereas earlier reports suggested a relative disconnect between ratings of insight and behavior (e.g. Ellis et al., 2011), subsequent research has provided evidence that subjective experience can in fact work effectively as a window into problem-solving performance. Crucially, Salvi et al. (2016) found that solutions rated as solved by insight tend to be more frequently correct than those solved by analysis. Webb et al., (2016) replicated such an effect. They also showed that problems designed to involve insight indeed yielded a stronger experience of overall insight than did non-insight problems, designed to be solved primarily by analysis. Moreover, the strength of insight experience correlated with accuracy on insight problems, but not on non-insight problems.

Importantly, Webb and colleagues probed four types of specific ratings beyond the overall feeling of insight (i.e. beyond whether Aha! occurred or not): confidence, impasse, pleasure, and surprise. In the insight problems, the overall Aha! experience was positively related to all

specific ratings except for impasse. By contrast, in non-insight problems, confidence was related negatively, and the three remaining ratings were virtually unrelated. That fact suggests that the experience of insight may emerge from a high consistency between metacognitive (high confidence) and affective (high pleasure) evaluation of one's experience (with surprise sharing both metacognitive and affective characteristics).

Danek and Wiley (2017) systematically examined the structure of subjective experience during problem solving, and found that performance in insight problems was accurately predicted by two primarily¹ metacognitive dimensions – suddenness and certainty – as well as by two primarily affective dimensions: pleasure and relief. However, they did not contrast these results with data from analytic problems to see whether these ratings specifically predict insight problems, or correlate with correct solutions generally. Drażek et al. (2019) applied Danek and Wiley's four-dimensional scale of subjective experience, and found that only suddenness clearly distinguished insight from analytic problems (the former problems yielded a double mean rating, as compared to the latter problems). The three remaining ratings were comparable between the two problem categories (for a similar pattern of data, see Table 1 in Webb et al., 2016). However, a detailed comparison of subjective experience across the problems was a focus of neither Drażek and colleagues nor Webb and colleagues, so how precisely subjective ratings vary across various problem types remains to be established.

Overall, the process of insight problem solving is thought to reflect creative and flexible thinking, so it can substantially diverge from typical inductive reasoning, found in RAPM. That may result in interesting differences in metacognitive experience between these two problem categories. Thus, contrasting subjective experience in insight problems with that in various RAPM items can inform us on processes engaged throughout the RAPM test.

Study goals

This work examined the structured metacognitive ratings in the form of questionnaires that probed participants' metacognitive experiences (Flavell, 1979) referring to current, on-going cognitive processing and the resulting affective states. The study focused on the monitoring and the evaluative metacognitive functions (Schraw, 1998). Considering the monitoring function, participants were asked to self-monitor their own cognitive operations leading to the solution and remember them for later report. With regard to the evaluative function, participants were required to appraise the solution once it emerged.

The ratings were acquired on (a) the adopted problem solving strategy (either the analytic or intuitive approach to a problem), (b) the experienced suddenness of solution, (c) the perceived certainty of its correctness, and (d) the accompanying affective rating of pleasure. These ratings were analyzed in order to: (1) track potential changes in these ratings across the RAPM items of varying difficulty, as well as (2) between the correct and incorrect solutions; (3) compare the ratings for consecutive levels of difficulty with the corresponding ratings for the insight problems; and (4) investigate the pattern of correlations among ratings and their links to individual performance on RAPM. When defining these four goals, it was assumed that the underlying processing might differ across RAPM items, and such differences can be validly

¹ No rating purely reflects a single kind of state; however, suddenness and certainty likely reflect monitoring of primarily cognitive events, and pleasure and relief reflect primarily affective states.

reflected by metacognitive ratings. Since multidimensional scales of subjective experience have never been used in RAPM, this study had a potential to bring about novel knowledge on processing in RAPM. As RAPM has been a very popular intelligence test in both basic and applied research, and has underpinned many findings pertaining to the role of cognitive ability in human functioning in various situations, it is pivotal to understand what processes and abilities it actually measures.

Method

Overall, 20 selected RAPM problems and four math insight problems were presented on the computer, and for each problem, the response, its latency, and answers to eight questions marked on the Likert scale were recorded. Afterwards, four typical analytic problems, four classic insight problems, and four matchstick algebra problems were administered, to be compared with the computerized test. The study's procedure had been preregistered at https://osf.io/pqtza/?view_only=b97e88b6059343a69b27df6aba60d4c3. The preregistration materials describe the issues of sample size and power, as well as many technical details omitted in the main text.

Participants and procedure

A total of 119 participants (not solely students) were recruited via the Internet. Each person was paid 6 Euros in the local currency. The 19 people who failed to provide at least four correct responses, that is, respond above the random level of 2.5 [= 20 items/8 response options per item], were excluded. The final sample consisted of 100 people (73 females, all white, aged 18–32, mean age 22.6, median age 22, $SD = 2.94$). The study took place at a psychological laboratory of a university, in groups consisting of 5–9 people. The study lasted from 2.5 to 3 h, depending on a group's performance. Drinks and snacks were provided. The tasks were applied in a fixed order: the computerized test (RAPM and insight items); three WM tasks (Paper Folding, Spatial Span, and Memory Updating); as well as the interactive insight problems and the paper-and-pencil analytic problems (see below). Due to procedural issues, the last five people omitted the interactive and paper-and-pencil problems. WM was measured to control for individual cognitive capacity, but it predicted subjective ratings similarly as did accuracy on RAPM, so these results were not reported.

Tasks

Computerized Raven advanced progressive matrices

Basing on our previous administration of 36 RAPM items in a sample of 172 young adults with a comparable procedure (randomized order, delayed response bank), for this study we selected 20 items that yielded accuracy between 80% and 20%, avoiding both trivial and virtually unsolvable items. The following matrices were selected: #8, 10, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 33, 34, and 35. The item sequence was fully randomized for each person. Unlike in paper-and-pencil RAPM, in which the matrix and the responses are presented on the same sheet, here, each item was presented in two steps. First, solely the matrix

was presented for 90 s (seconds) or until the button “ready for responding” was pressed. Then, the response bank was presented for maximum 15 s. The total of 105 s per item constituted a relatively low time pressure, allowing the participants to both solve the RAPM and monitor their mental states (originally, 2400 s is recommended for 36 problems, i.e. 67 s per problem on average). The response was selected by clicking with the mouse.

Computerized insight problems

Four math problems were used previously in insight research (e.g. included in the Cognitive Reflection Test; Frederick, 2005; Toplak et al., 2011). Each problem was presented at the center of the screen for 90 s, and then eight responses were presented for 15 s, matching closely the procedure for RAPM. In our previous studies, average accuracy for these insight problems oscillated around 50% – a value corresponding to accuracy on RAPM.

The problems were the following (with responses, the correct one bolded):

1. 10 machines make 10 toys in 10 min. How much time is needed for 100 such machines to make 100 toys? (1, 2, 5, **10**, 50, 100, 1000, 10,000 min)
2. John drinks 1 crate of beer in 6 days. Mary drinks 1 crate of beer in 12 days. In how many days they will drink one and the same crate of beer? (1, 2, 3, **4**, 6, 9, 10, 12 days)
3. You bought a book for 60\$ and then sold it for 70\$. Then you bought it again for 80\$ and sold it for 90\$. How much did you earn? (10, **20**, 30, 40, 50, 60, 70, 80, 90\$)
4. In the family there are 7 sisters. The 3 youngest are triplets, and the 2 oldest are twins. Each sister has only 1 brother. How many brothers are in the family? (**1**, 2, 3, 4, 5, 6, 7, 8 brothers)

Each of the insight problems was placed randomly in the sequence of RAPM items. The sequence was preceded by two training RAPM items (#3 and #6) and by the giraffe-and-ostriches insight problem. Additionally, research assistants explained the task.

Non-computerized insight problems

However, for most of the insight problems, the solution cannot be shown among other options, because it is immediately grasped as correct. Therefore, such items could not be included in the computerized procedure. Thus, another four typical insight problems (Lilies, Bat & Ball, Socks, Eight Coins; see Weisberg, 1995) and four widely used matchstick algebra problems ($I=II-II$, $VI=VI+VI$, $I-I=I$, $IV=IV+V$; see Knöblich et al., 1999) were applied after the computerized test. Lilies and Bat & Ball were reformulated (with the structure intact), as they became too familiar. The Bat & Ball variant was: “A flashlight and a battery cost 11\$ in total. The flashlight costs 10\$ more than the battery. How much does the flashlight cost?” The Lilies variant was: “Bacteria doubles in area in a pot every 24 hours. There was only one bacteria 24 days ago. Today, bacteria covers the entire pot. How many days ago did it cover the half of pot?”. All the problems were presented in an interactive format, that is, small balls representing bacteria, real socks, actual coins, and plastic sticks were provided, to increase the chance for genuine insight (see Weller et al., 2011). Four minutes were allowed for each matchstick problem, and 2 min were given for each classic insight problem (the times were based on previous research on interactive insight problems; see Chuderski et al., 2020; Weller et al., 2011).

Non-computerized analytic problems

Four typical analytic problems (Crime, Grandmother, Bachelor, Languages) that require complex combinatorics were administered in the paper-and-pencil format (with responses to be written down). For instance, the Crime problem stated: “Either A, B, C, or D committed a crime. Each of them made a statement, but only one statement was true. A said: ‘I didn’t do it.’ B said: ‘A is lying.’ C said: ‘B is lying.’ D said: ‘B did it.’ Who committed the crime?” 4 min were given to each problem, to match with the matchstick problems.

Subjective experience scale

After selecting each response in the computerized task, eight questions probed subjective experience on the Likert-like scale ranging from 1 to 19, at which the participants located her or his particular experience between the two extreme states:

1. I solved the problem: analyzing consecutive elements vs. looking at the entire problem;
2. I found the problem solution: intuitively providing an answer vs. systematically discovering particular rules;
3. The solution in my head: was gradually emerging vs. appeared suddenly;
4. The appearance of the solution in my head was: totally unexpected vs. fully expected;
5. After the solution, my feelings were: unpleasant vs. pleasant;
6. Finding the solution elicited my: satisfaction vs. discontent;
7. My feeling regarding the correctness of response was: uncertainty vs. certainty;
8. I selected the response: decisively vs. hesitantly.

The middle point of each scale was marked with “in between.” The instruction was: “Please describe your subjective experience at the moment when you found the solution to this problem.” Scales were formulated in the local language. All questions needed to be answered to move to the next problem. The questions in fact probed four dimensions (ratings), with even-numbered questions comprising reversals of the preceding odd-numbered questions (to control for the left–right-side response bias). The four ratings were: Intuition, Suddenness, Pleasure, and Certainty. The first rating was loosely modelled after the strategy questionnaire used in Gonthier and Tomassin (2015). The other three ratings were modelled after Danek and Wiley (2017; their Relief option was omitted, as our previous data suggested that it strongly overlapped with Pleasure, $r = .516$). At the beginning, the participants read instructions on how to monitor their internal experience accompanying problem solutions.

For both the non-computerized insight and the analytic problems, questions followed solely the correct solutions, because in these problems the incorrect solutions are provided extremely rarely – typically, on the failed attempts, no solution is provided at all. These two types of problems acted only to validate the subjective ratings. As a result, they could define the upper and the lower limits for the subjective experience: either a highly intuitive, sudden, pleasant, and confident experience; or an analytic, systematic, potentially frustrating, and questioning experience, respectively. Ratings from RAPM could then be contrasted.

Dependent variables

The main dependent variables for both the computerized and non-computerized tasks included the four subjective ratings. Three ratings were calculated as the Likert score in the odd-numbered question + (20 – the score in the even-numbered question), divided by 2. One exception was question #4, which appeared to be understood by the participants differently than originally intended. “Fully expected” was intended to mean the opposite to “appeared suddenly,” but in fact the two statements correlated positively. As it came up, solutions judged as appearing the most suddenly occurred in the easiest RAPM items, that is, in which correct solutions were frequent. Therefore, a participant could easily expect that the solution would occur, only they could not predict exactly *when*. Ergo, in easy items, the high values were rated both on expectedness and on suddenness (and vice versa in difficult items). To cope with this procedural error, we chose to use only question #3 as the rating for Suddenness, and nevertheless, it yielded clear effects. Overall, the internal consistency of each rating across items was high (Cronbach's alpha ranging from .77 to .86).

Additionally, for the computerized problems, the mean solution time (ST) was analyzed (the time elapsing from the presentation of the problem until the “ready for responding” button was pressed). For the paper-and-pencil problems, no STs were collected.

Hypotheses

The following four hypotheses were formulated with regard to the subjective ratings.

- H1: the effect of RAPM item difficulty on the ratings. Subjective experience varies reliably between the easier and the more difficult RAPM items. Two alternative patterns are possible. H1a: More accessible solutions can be reached in a well-defined sequence of elementary cognitive operations (attention, memory, imagery, reasoning, etc.), while discovering more demanding solutions that require more than just such well-defined sequences may lead to experiencing insight-like states; thus, subjective ratings (all or some of them) increase with increasing difficulty. Possibly, the ratings for the easiest items fall reliably below the criterion rating 10 (= “in between” on the scale), suggesting an analytic process, while the hardest items surpass 10, suggesting insight. H1b: Alternatively, experiencing insight-like states might result from fluency of processing (Topolinsky & Reber, 2010) – likely in the easier items. The more demanding the RAPM items become, requiring more complex and intensive combinations of problem elements, the more effortful and equivocal feelings might be experienced; thus, the subjective ratings (all or some of them) decrease with increasing difficulty. Possibly, the ratings for the easiest items reliably surpass the score of 10, while the hardest items might fall below this score.
- H2: the effect of solution's correctness on the ratings. Assuming H1b, the ratings for the correct and incorrect solutions in RAPM might differ but also interact. For easier items, the correct solutions might yield relatively higher ratings than the incorrect solutions, indicating relatively effortless and fluent integration of the governing rules when the correct response was selected, with errors driven by more effortful and less fluent processing in cases when such an integration could not be achieved. For the more difficult items, the incorrect solutions might yield relatively similar ratings as those accompanying the correct solutions, indicating effortful reasoning processes in both of them.

- H3: the comparison of the ratings in RAPM vs. insight items. Assuming H1b, easier Raven items yield a subjective experience more similar to that of insight problems than the experience yielded by more difficult items, but still the ratings obtained from RAPM might be substantially lower, as compared to those in insight problems.
- H4: correlations amongst the ratings in RAPM vs. insight items as well as the ratings' relationships with individual RAPM scores. Negligible inter-correlations amongst the subjective ratings are predicted for the ratings reported for correct solutions in RAPM, matching previously reported data on the null relationships among the ratings for analytic problems (Webb et al., 2016). Moderate to strong inter-correlations of particular ratings are predicted for the insight items, as observed in previous reports (Danek & Wiley, 2017; Webb et al., 2016). Individual RAPM scores might be negatively predicted by average subjective ratings (higher ratings will accompany lower RAPM scores), with an exception for Certainty (people higher in fluid intelligence are expected to respond in a more confident way).

Data screening and analysis

As many as 145 RAPM trials (7.2%) were excluded from the analysis of ratings because either $ST < 10$ s (suggesting a random guess; inspecting the matrix and inferring the solution within 10 s was unlikely), or 90 s elapsed with no option selected (the subsequent questions had no reference). Eleven insight trials (2.7%) were excluded for the same reasons.

Data were analyzed using Bayesian t test, BANOVA (F test), and Bayesian Pearson correlation (r). The Bayesian variants of analyses yield two straightforward advantages over the frequentist variants that compute p values. First, Bayesian variants use explicitly the probability distributions (and not only means and standard errors) to draw inferences about the plausibility of a hypothesis. Second, they assess precisely the relative strength of evidence for the null and for the alternative hypothesis (and not only test the alternative hypothesis probability, while assuming that the null is true). Specifically, the Bayes factor (BF_{10}) reflects the ratio of the likelihood of the data under the model assuming an effect (difference, correlation, etc.) to the likelihood of the data under the model lacking such an effect (for an accessible guide to Bayesian statistics see Stern, 2016). This allows to assess not only evidence in favor of the hypothesis, but also evidence against it. Values of BF_{10} above 3.0 (or 0.33) were interpreted as suggesting that evidence supports the hypothesis assuming the effect (or rejects it), with $BF_{10} > 10$ suggesting strong support. BF_{10} values between 0.33 and 3.0 were interpreted as inconclusive (the criteria adopted after Kass & Raftery, 1995).

All analyses were performed using JASP 0.9.1 (jasp-stats.org). As the present study was the first one to measure the four-dimensional subjective experience in RAPM, and no precise expectations could be formulated about the size of potential effects, the default JASP priors were used. Regarding the fixed effects in the t test, for H0 the spike prior pointed at zero, and for H1 a Cauchy distribution was centered on zero with scale parameter equal to .707, indicating that the expected median of Cohen's d for the effect was .707. For the BANOVA, the scale parameter equaled 0.5 for fixed effects and 1.0 for random effects. For the correlation, the stretched beta prior width equaled 1.0, that is, all correlation strengths between $r = -1.0$ and $r = 1.0$ were expected as equally likely (see Wagenmakers et al., 2018).

Results

Comparing subjective reports for the non-computerized problems

Before testing the four hypotheses, the validity of consecutive ratings was assessed by comparing their mean values between the 244 interactive insight problems solved and 168 paper-and-pencil analytic problems solved. Figure 2 presents these mean subjective ratings and the respective BF values. Intuition did validly differentiate the problem categories, reliably exceeding the criterion value 10 for the insight problems, $BF_{10} = 86.07$, but falling below 10 for the analytic problems, $BF_{10} > 999$. Additionally, Suddenness distinguished the categories, reliably exceeding 10 for the insight problems, $BF_{10} > 999$, while falling below 10 for the analytic problems, $BF_{10} > 999$. By contrast, Pleasure did not differentiate the problems, and its values always exceeded 10, $BF_{10} > 999$. Although $BF_{10} = 2.54$ for Certainty did not indicate reliable evidence, there was some trend for the insight problem solutions to yield larger confidence. As the overall number of correct solutions to typical problems was relatively low, it was expected that this rating might yield more conclusive results in larger pools of items (e.g. for RAPM; see below). Crucially, Pleasure strongly correlated with Certainty, both for the insight problems, $r = .712$, and the analytic problems, $r = .782$, each $BF_{10} > 999$, while all other correlations were much weaker, with r s = .007 to .325. Therefore, Pleasure and Certainty contained highly overlapping information, and it seems that Pleasure likely indexed simply a positive affect related with finding any problem solution with a given confidence. This conjecture was confirmed by checking the RAPM data, which yielded an even stronger correlation between the two ratings, $r = .820$. Thus, to simplify further analyses, only Certainty was taken into account, whereas the redundant data for Pleasure were skipped.

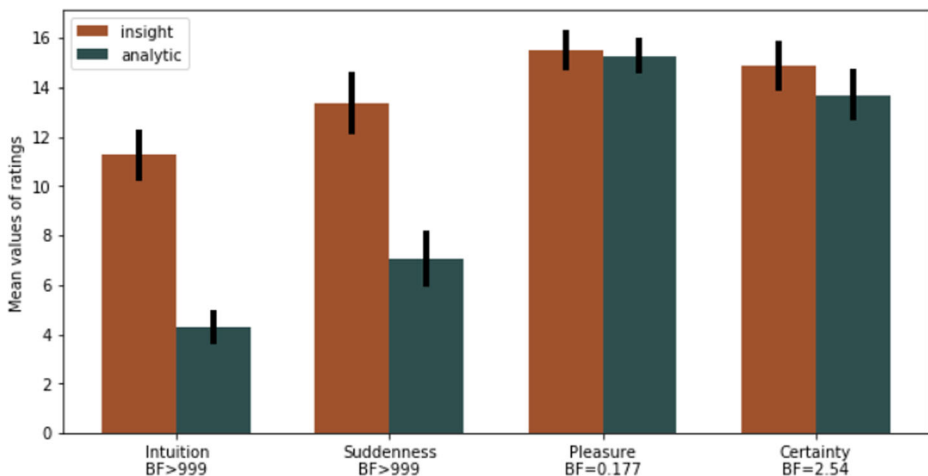


Fig. 2 Mean values of subjective ratings (note that their full scale ranged 1–19) following the 244 correct solutions on non-computerized insight problems vs. the 168 correct solutions on the paper-and-pencil analytic problems; the BF values assess evidence that a given rating on insight problems exceeds the analytic problems; bars represent 95% credible intervals

Overall performance in computerized problems

Mean accuracy on the 20 RAPM items equaled $M = .51$ (participants' range .20–.85). The lowest error rate was noted for items #16 ($M = .10$) and #14 ($M = .22$); it increased up to items #34 ($M = .80$) and #33 ($M = .83$). Accuracy in the insight items equaled $M = .55$ (range 0–1.0), matching the mean accuracy on RAPM. ST increased between the fastest RAPM item #10 ($M = 20.5$ s) and the slowest item #28 ($M = 67.7$ s), $BF_{10} > 999$. Correct solutions were delivered faster ($M = 37.7$ s) than incorrect ones ($M = 45.4$ s), $BF_{10} > 999$. Mean ST in insight items equaled $M = 40.7$ s, and did not differ between correct and incorrect trials.

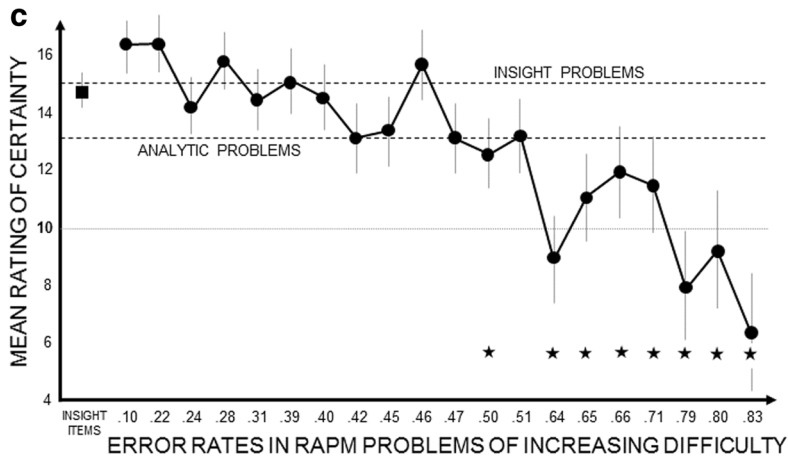
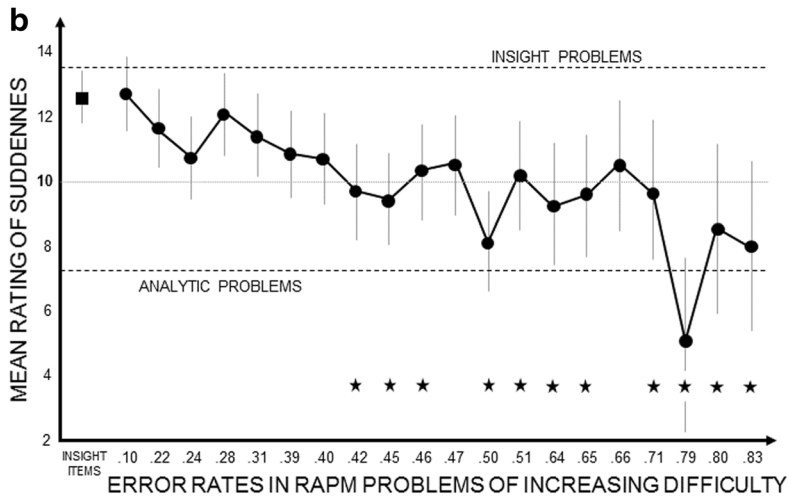
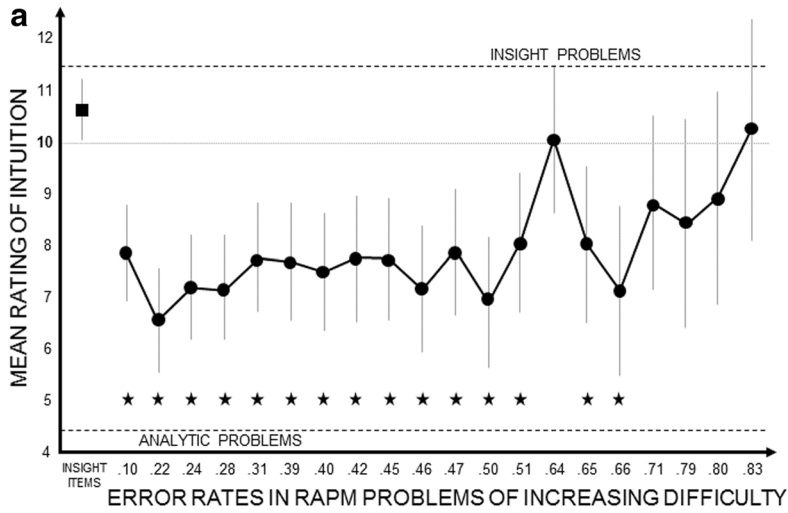
Subjective reports in computerized problems as a function of difficulty (H1)

To analyze the effect of item difficulty on the three ratings (Intuition, Suddenness, Certainty) that followed correct solutions, 1017 values of each rating were submitted to BANOVA, with the RAPM item as a fixed factor, and participant as a random factor (i.e. for each participant, her/his mean rating value was separately accounted for by the model). Intuition did not vary reliably across the 20 RAPM items, with $BF_{10} = 0.060$, supporting the null effect. As can be seen in Fig. 3a, the upper bounds of the rating's 95% credible intervals excluded the criterion value 10 for the majority of items except the few most difficult items (pronounced intervals for these latter items resulted from infrequent correct responses to them), suggesting that generally the participants adopted the strategies that relied on analytic and not intuitive processing. The lack of reliable item effect was also indicated by small effect size, $\eta_p^2 = .03$. In contrast, Suddenness reliably decreased as a function of the item difficulty, $BF_{10} > 999$, and yielded a moderate effect size of $\eta_p^2 = .07$. For several easier items, its lower 95% credible intervals excluded 10 (Fig. 3b), suggesting a sudden emergence of the solution, while the values for virtually all other items oscillated around 10, with two values reliably below 10. A reliable effect was observed also for Certainty, $BF_{10} > 999$, and it was strong, $\eta_p^2 = .24$ (Pleasure yielded a virtually identical effect). This time, 13 easier items yielded the 95% credible intervals above 10, and the remaining items' ratings oscillated around 10, with the most difficult item located reliably below 10 (Fig. 3c).

Subjective ratings for the correct versus incorrect solutions (H2)

To examine the effect of solution correctness on ratings, the BANOVA model was run on all the 1855 values of each rating (for 1017 correct and 838 incorrect solutions), with RAPM item and solution correctness as two fixed factors, and participant as a random factor. The analysis indicated that for each rating there was a reliable effect of correctness, each $BF_{10} > 999$, $\eta_p^2 = .02$ for Intuition, $\eta_p^2 = .05$ for Suddenness, and $\eta^2 = .21$ for Certainty. The correct trials yielded larger values of Suddenness ($\Delta M = 2.52$) and Certainty ($\Delta M = 6.36$), but were accompanied by lower values of Intuition ($\Delta M = -1.32$), as compared to the incorrect trials. For Intuition and Suddenness, there was no reliable interaction of item and correctness, $BF_{10} =$

Fig. 3 Mean values of the subjective rating of Intuition (a), Suddenness (b), and Certainty (c) for the 20 RAPM problems (dots) and the average on 4 insight problems (squares); stars indicate the RAPM ratings that are reliably lower than the insight rating ($BF_{10} > 3.0$); dashed lines reflect mean rating in typical insight problems (in interactive format) and paper-and-pencil analytic problems; bars represent 95% credible intervals; full rating scale ranged 1–19



.006 and $BF_{10} = .084$, respectively, suggesting that item difficulty did not affect the difference in rating between the correct and incorrect trials. In contrast, for Certainty there was a reliable interaction, $BF_{10} = 230.7$. It indicated a gradual decrease of the difference in rating between the correct and incorrect trials. In the five easiest items this difference equaled $\Delta M = 5.94$, whereas in the five hardest items it dropped to $\Delta M = 3.88$. However, the latter difference was still highly reliable, $BF_{10} > 999$.

Subjective ratings in RAPM versus insight problems (H3)

The effect of correctness on Intuition was the opposite for the computerized insight problems, as compared to RAPM. The correct trials yielded higher ratings (instead of lower ratings in RAPM), $\Delta M = 1.41$, although the correct-incorrect difference only approached the assumed level of effect reliability, $BF_{10} = 2.81$. However, the task (RAPM vs. insight) \times correctness (correct vs. incorrect) interaction was highly reliable, $BF_{10} > 999$. By contrast, the direction of difference for Suddenness and Certainty matched the RAPM data, with correct trials related with larger values of the two ratings, $\Delta M = 2.50$, $BF_{10} = 119.2$, and $\Delta M = 3.35$, $BF_{10} > 999$, respectively.

In Fig. 3a–c, the mean ratings and their 95% credible intervals for the correct trials of computerized insight problems are presented at the far left. The RAPM items that yielded reliably lower values of a given rating than the insight problems (as indicated by $BF_{10} > 3.0$) were marked with a star. For Intuition, 15 RAPM items yielded reliably lower ratings than the insight items' rating, but five comparisons (for difficult items) were inconclusive. For Suddenness, as many as 11 moderate and difficult RAPM items yielded reliably lower ratings than did insight items, the 8 easy items yielded inconclusive comparisons, and the easiest item even surpassed the insight items value. For Certainty, 8 difficult RAPM item ratings were reliably lower than for the rating for insight items, 8 primarily easy items surpassed or equated it, and 4 moderate items yielded an inconclusive difference.

Finally, subjective experience in the correct RAPM solutions could be mapped onto experience in the typical insight and the typical analytic problems. Figures 3a–c indicate that RAPM involved reliably higher Intuition than the analytic problems, but reliably lower Intuition than reported in the non-computerized insight problems (except for two RAPM items). RAPM solutions occurred as less sudden than the solutions in the insight problems (except for the easiest item), but also less gradual than in the analytic problems (except for the three most difficult items). The two easiest items yielded Certainty that was even higher than in the insight problems, while several most difficult items involved lower Certainty than in the analytic problems. Thus, it seems that although solving RAPM items was steadily related with a mix (or an intermediate) of strategies applied in the insight and in the analytic problems (little variance in strategy across items), the time course and robustness of a solution spanned substantially, from the high values typical for insight problems (for the easier items), down to the low values characteristic for the analytic problems (for the harder items).

Correlations among subjective ratings and performance (H4)

Correlations for ST and the three ratings were comparable between the correct and the incorrect RAPM solutions: All correlations were reliable in both matrices, each $BF_{01} > 999$ (except for the Intuition–ST link, $r = .043$, $BF_{01} = 0.17$), the corresponding r values had the same sign, and the mean absolute difference between them equaled only $\Delta r = .06$. Therefore,

correlations were calculated across all the 1855 solutions. Suddenness correlated positively with Certainty, $r = .403$, and Intuition, $r = .188$. By contrast, Certainty and Intuition correlated negatively, $r = -.320$. A shorter ST was associated with larger Suddenness, $r = -.282$, and Certainty, $r = -.354$. Across the 389 insight items, Suddenness also correlated positively with Certainty, $r = .461$, and Intuition, $r = .374$. One difference to RAPM was that the negative correlation of Certainty and Intuition, $r = -.039$, was not reliable. Each rating negatively correlated with ST, $r = -.162$, $r = -.463$, and $r = -.458$, respectively. Across 100 participants, the RAPM score correlated reliably only with Certainty, $r = .328$, $BF_{10} = 30.55$. RAPM's links with Intuition, Suddenness, and ST were not reliable, each $r < .15$, each $BF_{10} < 0.35$.

Discussion

This study aimed to extend our understanding of the mechanisms that are crucial for solving a hallmark fluid intelligence test (RAPM) by observing how metacognitive ratings change across 20 RAPM items of varying difficulty, and comparing them to 4 interleaved insight problems. The response bank was shown late, promoting mental construction of the solutions, and discouraging from using other, more simplified strategies (such as response elimination) known to decrease the RAPM validity. After each response, the participant rated her or his subjective experience of how intuitive and holistic her or his problem solving strategy was (as opposed to systematic and analytic); how suddenly the solution emerged in the mind (as opposed to emerging gradually); how much satisfaction it yielded (as opposed to discontent); and how certain and decisive its choice was (as opposed to uncertain and hesitant). High values of these four ratings are typically associated with experiencing insight, or Aha! moments.

These ratings were first validated using classic insight problems and typical analytic problems – two relatively well understood kinds of problems. It appeared that, indeed, insight problems involved relatively larger values of the ratings, while analytic problems yielded lower values (except for Pleasure, which additionally highly overlapped with Certainty).

Assuming that the ratings (all or some of them) vary across the RAPM items (H1), it was tested whether they either increase with increasing difficulty (H1a) or decrease (H1b), depending on whether subjective experience typical for insight results either from the non-obvious discovery of hidden complex rules in difficult RAPM items or the fluent processing of simple rules in easy items, respectively. It was also examined whether any rating can be larger during the correct solutions than during the incorrect solutions (H2). Next, it was predicted that even if H1 is confirmed, all the ratings for RAPM items may still be lower than the ratings in computerized insight problems (H3). Finally, it was expected that the ratings, which typically positively correlate in the case of insight problems, may be less correlated for the RAPM items and, with the exception of Certainty, may be negatively related with performance on RAPM (H4).

Regarding the first question, hypothesis H1b was supported, and H1a was not. Suddenness as well as Certainty (and the highly overlapping Pleasure) decreased from the easier to more difficult RAPM items, from the levels indicating insight (ratings above the criterion value 10 and indifferent from the insight items), to values typical for analytic processing (ratings below 10 and thus much below the corresponding values for the insight items). Therefore, this pattern seems to match Topolinsky and Reber's (2010) interpretation that subjective experience of sudden and vivid comprehension of the problem solution results from the momentary boost in processing fluency. High suddenness and certainty of the correct solution (as well as pleasure

from it) were reported for the RAPM items that were relatively easy and fast, implicating high fluency leading to this kind of insight-like experience. In contrast, correct solutions to the slow and difficult RAPM items emerged gradually, involving low certainty (and low pleasure). Thus, solutions for difficult RAPM items might not elicit a sudden boost in processing fluency leading to a discovery of a hidden rule (i.e. from change in the problem representation), but might be delivered by means of the systematic, effortful, and complex (and thus less certain) combination of the consecutive elements of the problem.

Only Intuition did not support H1b (and H1a), as it was not affected by the RAPM item difficulty. It might be speculated that the null effect might have resulted from delayed presentation of the response bank, which might have forced participants to use the same strategy on all the items. Relatively low and constant values of the rating (around 8) suggest that the participants relied on a strategy that combined systematic, analytic processing with a little bit of intuitive processing, as compared to the fully systematic strategy reported in the analytic problems (the rating close to 4). In the typical RAPM (with concurrent presentation of problems and response options), the effect of difficulty on strategy could be more evident; however, such a presentation taps into fluid intelligence less validly (Becker et al., 2016).

The correct solutions were related with lower Intuition and higher Certainty, which confirms that systematic analysis helps in RAPM (see Gonthier & Tomassin, 2015; Hayes et al., 2011; Jarosz et al., 2019; Jarosz & Wiley, 2012; Jastrzębski et al., 2018; Loesche et al., 2015; Mitchum & Kelley, 2010; Vigneau et al., 2006). That correct solutions emerged more suddenly than incorrect ones might have resulted from the fact that in many incorrect trials, participants reached no solution at all (not even an incorrect one), so they had to select an arbitrary response; such a choice is unlikely to be experienced as sudden. Only Certainty involved a reliable interaction of RAPM item and correctness, with a decreasing gap in the rating values between the correct and incorrect solutions as the item difficulty increased. Ergo, H2 seems to be generally supported, apart from that for Intuition and Suddenness the difference between the correct and incorrect solutions was not affected by item difficulty.

An interesting result concerns the reliable interaction of correctness and problem type (RAPM vs. insight) on the strategy reported. In contrast to RAPM, which clearly benefited from more analytic strategy, the insight items tended to benefit from more intuitive strategy. However, the latter effect was not robust enough to claim that analytic and systematic processing was detrimental to the four computerized insight problems.

The test of H3 yielded the most intriguing results. Although for most of the RAPM items, Suddenness and Certainty were much lower than the values for insight items (i.e. the values were more typical to analytic problems), confirming H3, subjective experience on the easiest RAPM items matched experience on the insight items – and even the classic problems presented in an interactive format (magnifying insight). This result additionally supported the view that under low RAPM complexity, the processing is highly fluent and univocal, unlike in the difficult problems. Intuition was reliably lower for virtually all RAPM items, with only two difficult items yielding experience that reached the criterion value 10 and approached the level observed for computerized insight items (but not for interactive problems).

Finally, hypothesis H4 was partially supported, but only with regard to correlations across participants. As predicted, Certainty correlated positively with the RAPM score. However, the latter score was unrelated with Intuition and Suddenness (whereas negative relationships were predicted). Specifically, the lack of negative link from Intuition to RAPM is surprising, as previous reports indicated that heuristic strategies relate negatively with fluid intelligence and

WMC (Chuderski & Jastrzębski, 2018; Gonthier & Tomassin, 2015; Jarosz & Wiley, 2012). Predictions about the null correlations across the RAPM solutions were disproved, as the three ratings inter-correlated reliably. Finally, more intuitive solutions to RAPM were also less certain, but no such relationship occurred in insight items.

Overall, data generally confirmed H1 (metacognitive experience, but not the adopted strategy, differs across the RAPM items of varying difficulty). H2 was also confirmed (correct solutions yield larger values of metacognitive ratings), but the advantage for the correct over incorrect solutions was general: The item factor only slightly affected this advantage in the case of Certainty. Evidence related to H3 was more complicated: Although coping with moderate and difficult RAPM items led to an experience resembling combinatorial, analytic processing, several of the easiest items elicited as sudden and as certain solutions as those reported for insight items and insight problems, suggesting a comparable processing fluency in all these items. For H4, contrary to our predictions, the correlations among ratings in RAPM solutions were reliable, and comparable with those for insight items (except for the Intuition–Certainty link). Across participants, the tendency towards an intuitive solution strategy was not detrimental to performance (the same held for the tendency towards suddenness). The sole correlational prediction confirmed was that people higher in RAPM scores report higher Certainty, suggesting that people who run more effective reasoning processes are also more confident in the results of these processes.

Therefore, what can the metacognitive ratings tell us about processing in RAPM, at least in a variant in which the matrix is temporally separated from the response bank? The reliable and meaningful differences in subjectively experienced processing observed between easy and difficult items (as well as large differences in solution times) suggest that RAPM consists of two kinds of problems. The easy problems involve concurrent and fluent processing, leading to insight-like, clear, and vivid outcomes (low complexity of these problems probably allows individuals to grasp most of the information at once and with high certainty). The difficult problems are too complex to grasp within one step, so they need to be decomposed into sub-problems and require systematic analysis of the consecutive problem elements, which ultimately must be integrated in an effortful and uncertain way. The two contrasting patterns of subjective experience, related to each kind of problem, do not depend on variation in strategy, because how much intuitive and holistic processes were relied upon was stable across RAPM items. Intuition was relatively low: The participants instead “analyzed consecutive elements” and “systematically discovered particular rules,” although on average, their strategy was a bit more intuitive than the full-blown systematic strategy reported for the analytic problems.

In consequence, the present study contrasts classic psychometric analyses suggesting unidimensionality of RAPM (e.g. Alderton & Larson, 1990; Arthur & Woehr, 1993), RAPM may not comprise a fully homogenous test taping solely reasoning ability, but the items' position and difficulty (both strongly correlating in the original paper-and-pencil variant) may contribute to additional variance that may constitute another dimension (see also Lozano, 2015; Ren et al., 2014).

Two exploratory analyses support the suggestion that early and late RAPM items seem to differ qualitatively. First, a reliable difference was found between a negative correlation of ST with the total RAPM score on the easiest five items, $r = -.126$, and its positive correlation on the hardest five items, $r = .377$, $BF_{10} > 999$, meaning that effective performance indicated rapidness on the easy problems, but deliberation on the difficult problems (see also Neubauer, 1990). Second, the mean values of Suddenness and Certainty (averaged over 100 participants) for the 20 RAPM items were almost perfectly predicted by the mean ST values for these items,

$r = -.889$ for Suddenness, $r = -.886$ for Certainty, each $BF_{10} > 999$. Thus, the shorter the time spent on a problem, generally the more insight-like the experience; the longer the time needed to solve a problem, the more systematic and reliable solution process was reported by the participants. In contrast to Suddenness and Certainty, the respective ST correlation with Intuition was positive, $r = .763$, suggesting that this rating did not index processing fluency.

One potential methodological consequence of the knowledge on RAPM gathered in this study is that time pressure during RAPM administration (e.g. reducing the recommended time from 40 to 20 min.) may substantially alter the validity of a resulting score. Under time pressure, this score may be loaded primarily by easier items (most people do not have time to attempt the most difficult items; see Estrada et al., 2017), and thus the score may reflect in the first place the fluency of “on-line” processing. With less time pressure, the scores may result primarily from individual differences in coping with more difficult items (with ample time, most participants will solve most of the easier items), and thus may reflect the capacity for systematic, combinatorial dealing with complex information.

The study also contributed to the understanding of subjective ratings, so far examined primarily for insight problem solving, by also applying them to RAPM. On the one hand, the ratings’ high sensitivity to the RAPM items of varied difficulty and latency, as well as clear differences detected between certain RAPM and insight problems, suggest high overall validity of the participants’ metacognitive ratings. On the other hand, the three ratings aimed to reflect three independent dimensions of metacognitive experience of insight (suddenness, pleasure, and certainty) were so strongly inter-correlated (which was especially true for the latter two ratings) – and they predicted solution time so closely – that their independence must be questioned, and they may index highly overlapping experiences instead (for a related conclusion, see Skaar & Reber, 2019). The present results partially match those of Webb et al. (2016), who also reported a strong link between pleasure and certainty (therein called confidence), but our results contrast with Webb et al.’s much less reliable links of these two ratings with surprise (perhaps the feeling of surprise accompanying the solution is not fully equivalent to experiencing its suddenness). Unlike the three ratings, the metacognition of the problem-solving strategy has been studied to date in inductive reasoning (i.e. outside insight problems). The present application of this kind of rating to the insight problems – revealing higher levels of intuitive processing involved in these problems – as compared to RAPM and analytic problems, supports the rating’s usefulness in research.

This study had some limitations. Subjective reports are by nature an imperfect method of accessing cognitive processes, and therefore our interpretation of the processing differences between easier vs. difficult RAPM items should be validated in future studies relying on objective methods. Even within the subjective domain, verbal protocols, less constrained than ratings, might provide a more precise (and less biased) window into processing in RAPM.

In conclusion, metacognitive ratings applied to Raven’s matrices – a popular test for intellectual assessment in basic, educational, and professional contexts – revealed that the subjectively perceived suddenness of a solution and confidence in its correctness differ between easy and difficult items, and therefore the test may be much more heterogeneous than commonly believed. Only the processing of difficult matrices is experienced as involving systematic and combinatorial decomposition and integration of matrix elements – that is, in line with the dominant models of the test (e.g. Carpenter et al., 1990) – while the easiest items involve more straightforward and fluent processing, typical for

insight problems. This novel knowledge can help in understanding the processes underlying solving the Raven matrices.

Acknowledgements

Data access Data can be freely downloaded from <https://osf.io/vr7zp/> or received upon request.

Funding information The study was supported by National Science Centre of Poland (project #2019/33/B/HS6/00321).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants Research was conducted on healthy, adult participants who volunteered for financial compensation, and were free to quit the experiment at any moment. The study did not include any controversial material. The study conformed to all ethical standards accepted in Poland for psychological research on healthy adults.

Informed consent Informed consent (signed in person) was received from each participant, according to the procedure accepted at Jagiellonian University in Krakow.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's advanced progressive matrices items. *Educational and Psychological Measurement*, 50, 887–900.
- Arendasy, M., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence*, 33, 307–324.
- Arthur, W., & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionality of the Raven's advanced progressive matrices. *Educational and Psychological Measurement*, 53, 471–478.
- Batchelder, W. H., & Alexander, G. E. (2012). Insight problem solving: A critical examination of the possibility of formal theory. *The Journal of Problem Solving*, 5, 6.
- Becker, N., Schmitz, F., Falk, A., Feldbrügge, J., Recktenwald, D., Wilhelm, O., Preckel, F., & Spinath, F. (2016). Preventing response elimination strategies improves the convergent validity of figural matrices. *Journal of Intelligence*, 4, 2.
- Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, 8, 205–238.
- 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, 404–431.
- Chu, Y., & MacGregor, J. N. (2011). Human performance on insight problem solving: A review. *The Journal of Problem Solving*, 3, 6.

- Chuderski, A., Jastrzębski, J., & Kucwaj, H. (2020). How physical interaction with insight problems affects solution rates, hint use, and cognitive load. *British Journal of Psychology*, Early View, <https://doi.org/10.1111/bjop.12442>
- Chuderski, A. & Jastrzębski, J. (2018). Much ado about Aha! Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General*, 147, 257–281.
- Colom, R., & Flores-Mendoza, C. E. (2007). Intelligence predicts scholastic achievement irrespective of SES factors: Evidence from Brazil. *Intelligence*, 35, 243–251.
- Cowan, N. (2016). Exploring the possible and necessary in working memory development. *Monographs of the Society for Research in Child Development*, 81, 149–158.
- Danek, A. H., & Wiley, J. (2017). What about false insights? Deconstructing the Aha! Experience along its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology*, 7, 2077.
- Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, 63, 453–482.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21.
- Demetriou, A., Makris, N., Tachmatzidis, D., Kazi, S., & Spanoudis, G. (2019). Decomposing the influence of mental processes on academic performance. *Intelligence*, 77, 101404.
- Drażnyk, D., Kumka, M., Zarzycka, K., Zguda, P., & Chuderski, A. (2019). No indication that the ego depletion manipulation can affect insight: A comment on DeCaro and Van Stockum. *Thinking & Reasoning*, 26, 1–33.
- Ellis, J. J., Glaholt, M. G., & Reingold, E. M. (2011). Eye movements reveal solution knowledge prior to insight. *Consciousness and Cognition*, 20, 768–776.
- Estrada, E., Román, F. J., Abad, F. J., & Colom, R. (2017). Separating power and speed components of standardized intelligence measures. *Intelligence*, 61, 159–168.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring. A new area of cognition-development inquiry. *American Psychologist*, 34, 906–911.
- Fleck, J. I., & Weisberg, R. W. (2004). The use of verbal protocols as data: An analysis of insight in the candle problem. *Memory & Cognition*, 32, 990–1006.
- Fleck, J. I., & Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology*, 25, 436–463.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.
- 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, 916–924.
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.
- 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.
- Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, 48, 1–14.
- Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence*, 40, 427–438.
- Jarosz, A. F., Raden, M. J., & Wiley, J. (2019). Working memory capacity and strategy use on the RAPM. *Intelligence*, 77, 101387.
- Jastrzębski, J., Ciechanowska, I., & Chuderski, A. (2018). The strong link between fluid intelligence and working memory cannot be explained away by strategy use. *Intelligence*, 66, 44–53.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Katona, G. (1940). *Organizing and memorizing studies in the psychology of learning and teaching*. New York: Columbia University Press.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1534–1555.
- Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, 65(1), 71–93.
- Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven advanced progressive matrices test. *Intelligence*, 48, 58–75.
- Lozano, J. H. (2015). Are impulsivity and intelligence truly related constructs? Evidence based on the fixed-links model. *Personality and Individual Differences*, 85, 192–198.
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10.

- Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's progressive matrices. *Intelligence*, 35, 359–368.
- Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 288–294.
- Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15, 238–246.
- 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, 699–710.
- Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence – their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131, 61–65.
- Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, 13, 179–212.
- Primi, R. (2001). Complexity of geometric inductive reasoning tasks. *Intelligence*, 30, 41–70.
- Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, 20, 446–451.
- Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. London: H. K. Lewis.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales. Section 3: Standard progressive matrices*. San Antonio: Harcourt.
- Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences*, 31, 30–35.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118–137.
- Salvi, C., Bricolo, E., Kounios, J., Bowden, E., & Beeman, M. (2016). Insight solutions are correct more often than analytic solutions. *Thinking & Reasoning*, 22, 443–460.
- Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122, 166–183.
- Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, 26, 113–125.
- Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, 72, 116–141.
- Skar, Ø. O., & Reber, R. (2019). The phenomenology of Aha-experiences. *Motivation Science*, 6, 49–60.
- Skrzypulec, B., & Chuderski, A. (2020). Nonlinear effects of spatial connectedness implicate hierarchically structured representations in visual working memory. *Journal of Memory and Language*, 113, 104124.
- Stern, H. S. (2016). A test by any other name: *P* values, Bayes factors, and statistical inference. *Multivariate Behavioral Research*, 51, 23–29.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39, 1275–1289.
- Topolinski, S., & Reber, R. (2010). Gaining insight into the “Aha” experience. *Current Directions in Psychological Science*, 19, 402–405.
- Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, 71, 1–26.
- Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, 34, 261–272.
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhaghen, J., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76.
- Webb, M. E., Little, D. R., & Cropper, S. J. (2016). Insight is not in the problem: Investigating insight in problem solving across task types. *Frontiers in Psychology*, 7, 1424.
- Weisberg, R. W. (1992). Metacognition and insight during problem solving: Comment on Metcalfe. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 426–431.
- Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 157–196). New York: Cambridge University Press.
- Weller, A., Villejoubert, G., & Vallée-Tourangeau, F. (2011). Interactive insight problem solving. *Thinking & Reasoning*, 17, 424–439.