Why People Fail on the Fluid Intelligence Tests

Article *in* Journal of Individual Differences · May 2015

DOI: 10.1027/1614-0001/a000164

Adam Chuderski
Jagiellonian University

Original Article

Why People Fail on the Fluid Intelligence Tests

Adam Chuderski

Jagiellonian University, Krakow, Poland

Abstract. The study examined the patterns of errors in a specially designed test of analogical reasoning. The results indicated that those patterns strongly depended on participants' ability level that was measured by another two fluid intelligence tests. Relatively good reasoners made analogy-making errors primarily resulting from not binding a single relational element to the complete solution. This fact indicates that they properly carried out a reasoning process, but missed just one reasoning step. In contrast, poor reasoners more often chose erroneous options that missed several relational elements, but were perceptually similar to target analogs, what suggests that those reasoners did not follow the necessary rules. Moreover, the reasoning scores of poor reasoners depended more strongly on measures of working memory capacity than did scores of good reasoners. The results are interpreted in terms of several seminal theories of fluid intelligence.

Keywords: fluid intelligence, reasoning ability, figural analogy, error patterns, working memory

Fluid intelligence (fluid reasoning, reasoning ability; Gf) is a human ability to use reasoning in order to solve novel abstract problems, which cannot be dealt with by making inferences from knowledge already possessed. Gf is usually assessed with matrix problems or visual analogy tests (Snow, Kyllonen, & Marshalek, 1984). One of the most important themes in fluid intelligence research consists of identifying the cognitive processes underlying fluid reasoning. So far, three main research methods were used in order to address this question.

Correlational Studies of Intelligence

The most common method to investigate factors that may determine the effectiveness of reasoning processes is the finding of elementary cognitive correlates of reasoning scores. Probably the strongest known predictor of fluid intelligence is working memory capacity (WMC), as it usually explains at least half of Gf variance (Kane, Hambrick, & Conway, 2005). However, as WMC measures are usually quite complex, their interpretation is an open issue, and several theories tried to explain what process can be responsible for WMC-Gf link. One influential theory assumes that individual performance in both WM tasks and Gf tests depends on the functioning of *executive control* exerted over cognitive processes, which includes directing attention, blocking distraction, and inhibiting responses (e.g., Burgess, Gray, Conway, & Braver, 2011; Kane, Conway, Hambrick, & Engle, 2007). The executive-control theory of reasoning explains that people with low attention control capabilities are poor reasoners because they suffer from poor maintenance of reasoning goals, as well as are prone to frequent capturing of their thinking by irrelevant stimuli/responses.

In contrast, it was shown that simple short-term memory (STM) tasks, which require little executive control, were at least as good predictors of Gf as were the variables representing executive control, once rehearsal and chunking were blocked in these tasks (Chuderski, Taraday, Necka, & Smolen, 2012; Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Cowan, Fristoe, Elliott, Brunner, & Saults, 2006). Thus, *storage capacity* (the number of distinct items simultaneously held in the active part of working memory) may primarily influence fluid reasoning, because it may determine how complex relations can be constructed from working memory items (e.g., Halford, Cowan, & Andrews, 2007; Oberauer, Süß, Wilhelm, & Sander, 2007). However, the debate between the executive-control and storage-capacity accounts of fluid reasoning is far from being settled.

Analysis of Error Options in Intelligence Tests

Important knowledge on processes that occur during fluid reasoning may also be provided by research relying on the application of specially designed fluid reasoning tests, and the detailed analysis of the types of errors that participants make. Especially, the changes in patterns of errors in relation to both ability level (i.e., whether good reasoners make different errors than poor reasoners) and manipulations applied (i.e., how a manipulation changes the observed patterns of errors) may bring crucial insights about the cognitive mechanisms underlying reasoning.

Journal of Individual Differences 2015; Vol. 36(3):138–149
DOI: 10.1027/1614-0001/a000164

The importance of error analysis was noted by the authors of the hallmark Gf test, Raven's Advanced Progressive Matrices (APM; Raven, Court, & Raven, 1983). Each item of this test includes a three-by-three matrix of figural patterns which is missing the bottom-right pattern, and eight response options are the patterns which could potentially match the missing one. The task is to discover the rules that govern the distribution of patterns, and to apply them to response options in order to choose the one and only right pattern. Raven et al. categorized incorrect response options into four categories. The most commonly made are so-called *incomplete-solution* errors, which are options which contain only a part of the correct solution, meaning that a participant most probably began reasoning effectively, but for some reason did not take into account all the rules that make the right choice. Also, there occur *overdetermined choices* – options which contain some features relevant for the correct solution, but also irrelevant ones. Less frequent are errors suggesting *an arbitrary line of reasoning* – the choice of an option that contains all the features represented in an item, but does not follow even one required rule. Also relatively rare, *repetition errors* are options identical to one of the three patterns adjacent to the missing pattern.

An interesting observation made by Forbes (1964) showed that a selected subsample of participants who scored 8 ± 1 in APM committed, in easy items that they could solve at all, as little as 23% of the incomplete-solution errors (out of all their errors). Two other subsamples that scored 18 ± 1 or 28 ± 1 in APM made, in more difficult items that they were able to solve, more than half of this type of error (52% or 58%, respectively). Although Forbes did not compare low- and high-ability people on the same items (nor on the whole test), his data suggested that both groups may differ in a way in which they solve the APM problems (for a similar conclusion see Babcock, 2002).

In a more recent study, Jarosz and Wiley (2012) reduced the number of response options in APM to four, and manipulated whether the most frequently chosen error option (a so-called salient distractor) was or was not a part of this set. They also tested whether salient distractors were either incomplete-solution or arbitrary-line-of-reasoning options. The modified test was applied to 64 students. The inclusion of salient distractors decreased reasoning accuracy, and increased its correlation with WMC. Items containing the arbitrary-line-of-reasoning options yielded a marginally stronger correlation with WMC than did items including the incomplete-solution options. This suggests that processing in problems in which the participants were far from the correct solution (because they primarily chose an arbitrary-line-of-reasoning option) depended on WMC more than did reasoning in the other type of problems, in which people were closer to the correct answer (i.e., they missed only few elements from it). Moreover (Experiment 2), eye-tracking showed that WMC was inversely related to the time spent focusing on salient distractors, thus extending the data from Vigneau, Caissie, and Bors (2006), who showed that when dealing with the Gf test, low-ability people spent most of their time inspecting response options, whereas high-ability ones devoted most of their time to matrix analysis (also see Bethell-Fox, Lohman, & Snow, 1984).

One disadvantage of the above taxonomy of errors is that it was constructed after the APM test had been designed, and that error options do not follow strict rules describing their deviation from the right option. Moreover, within an APM item there may be distinct numbers of errors from each category, which make their analysis difficult. In order to solve this problem, sometimes researchers construct their own intelligence tests, in which the number and internal structure of particular error options are controlled in a better way than in APM. In one such study, Vodegel Matzen, van der Molen, and Dudink (1994; Experiment 2) constructed a variant of Raven (equivalent to the standard version; SPM), in which the proper solutions contained either two or three rules, while error options omitted either one (single omissions) or more rules (multiple omissions). Vodegel Matzen et al. examined 200 children, and found that single omissions were 3.5 times more frequent than multiple omissions. However, as this pattern of errors was the same for low-, moderate-, and high-ability children, this result cannot explain why children differ in reasoning. One reason for the null effect observed might be that the SPM equivalent was too easy to solve to reveal any differences in the erroneous decisions of participants.

Computational Modeling of Reasoning in Intelligence Tests

The third method of investigating cognitive mechanisms of intelligence consists in computational modeling of the fluid reasoning process by using computational models and simulations (e.g., Evans, 1968; Kunda, McGreggor, & Goel, 2013; Lovett, Forbus, & Usher, 2010). The seminal computational model of APM has been developed by Carpenter, Just, and Shell (1990). The authors used eye tracking, on-line verbal protocols, and posttest verbal reports in order to analyze the line of reasoning that people applied to APM items. This data suggested that APM items are solved in an incremental way, as participants induced and applied rules governing the correct solution one by one. Moreover, the induction of each rule relied on the sequence of pairwise comparisons of adjacent elements of a matrix. Carpenter et al. fitted two variants of their APM model to performance indices of the average versus best reasoners. The average-ability variant committed errors in 11 problems out of 34 ones it was given. These errors (so, probably also errors of median scoring participants) primarily resulted from (a) inability to induce one specific rule (i.e., the distribution-of-two-values rule) as well as from (b) losing track of subproducts of a reasoning process because of insufficient WM resources. Thus, the average-ability variant often chose responses based on a very simplified (and thus incorrect) representation of a problem. In contrast, the high-ability variant (and probably the best human reasoners), which solved 32 problems, was able to simultaneously track multiple goals within its WM, and thus represented the complete (or almost complete) structure of relations defining a problem. Carpenter et al. supported the above conclusion by showing that scores in APM strongly correlated with performance on the Tower of Hanoi problems, which relied on effective goal manipulation in WM to a great extent.

Goals of the Present Study

Apparently, so far the studies using each of the above discussed methods were quite mutually isolated. Moreover, their results seem to be inconclusive to some extent. On one hand, within correlational studies on intelligence, there is no consensus on whether executive control or storage capacity is the primary underlying mechanism. On the other hand, the existing research on errors made in fluid reasoning tests, though inspiring, includes only a few studies, which due to the use of a simple version of such a test as well as the examination of young children sample (Vodegel Matzen et al., 1994), due to very limited adult samples (Jarosz & Wiley, 2012; Vigneau et al., 2006), or due to arbitrarily selected items/subsamples (Forbes, 1964), cannot be easily generalized.

Finally, the empirical results on WM-Gf correlations as well as error patterns in intelligence tests have not yet been interpreted in terms of existing computational models of reasoning in intelligence tests, which could provide us with a much better understanding of these results.

Especially, no study has yet showed, with the full and large sample of adult participants, that people (and corresponding variants of the APM model) supposed by Carpenter et al. (1990) to not follow the correct line of reasoning, differ in type of errors they commit, and get lower scores in the benchmark intelligence tests, in comparison to people who according to Carpenter et al. maintain in WM the (almost) complete relational representation of an attempted problem, and who may score higher on the same benchmark tests. Moreover, no existing study has yet investigated whether patterns of correlation between particular WM functions and reasoning performance of people not following the correct line of reasoning differ from the corresponding pattern of correlations regarding people who do follow such line of reasoning. Answering these two questions should allow for a better integration of knowledge gained from three so far isolated research methods, and it may also bring crucial insights into the nature of intelligence.

Consequently, the goal of the present study is to address these issues by investigating fluid reasoning within a novel figural analogy test (Figural ANalogies; FAN), in which (unlike in the existing Gf tests) the precisely defined types of error options were introduced, and their number was balanced within each test item. It was expected that people who differ in the patterns of errors in FAN would also differ in their scores on the benchmark reasoning tests. Specifically, it was predicted that people who more frequently choose incorrect options that are close to the correct solution (meaning that they undertook the proper reasoning process that failed at some stage), instead of options that are similar perceptually to target patterns, but far from the correct solution (meaning that people who choose these options used too simplified reasoning strategies or even did not undertake reasoning at all), will display higher level of fluid intelligence, as measured by the benchmark tests.

Moreover, it will be possible to examine whether in people differing in patterns of FAN errors the difference can also be found with regard to patterns of relationships between fluid intelligence and the elementary WM mechanisms involved in the maintenance of problem's internal representations. Specifically, the performance of people who end up their reasoning quite far from the correct solution (e.g., who choose primarily an arbitrary-line-of-reasoning option), and whose WM resources are possibly scarce, may be more dependent on WM capacity than the performance of people who may have enough WM resources to track the components of the correct solution (but somehow fail to bind them all), and who come closer to the correct solutions. This possibility has been predicted by Carpenter et al. model, as in the cases of failures its less effective version "tried to confirm all the rules in parallel, . . . but the resulting bookkeeping load was unmanageable" (p. 419). Due to a more effective control over its processing, the high-ability variant seemed to be less dependent on available capacity, as its version severely limited to just three rules that could be maintained in parallel solved a comparable number of problems as did the other variant without such a severe limit (21 vs. 23 problems, respectively). If the expected differences in the WM-Gf link are found, this fact will additionally support the Carpenter et al. model, and may shed light on the mechanisms responsible for the emergence of the strong relationship between WM and intelligence.

Method

Participants

Volunteer participants were recruited via publicly accessible social networking websites. Each participant gave informed consent and was paid the equivalent of 15€ in Polish zloty. A total of 347 people participated (210 women). The mean age was 24.3 years (*SD* = 5.03, range 18–45). All participants had normal or corrected-to-normal vision.

Materials

Computerized Figural Analogy Test

The FAN test consisted of 32 trials, presented serially on a computer screen. Each trial included two panels: one left and one right. The left panel contained three patterns arranged vertically, which formed the analogy "A is to B" as "C is to ?", where "?" was a missing pattern.

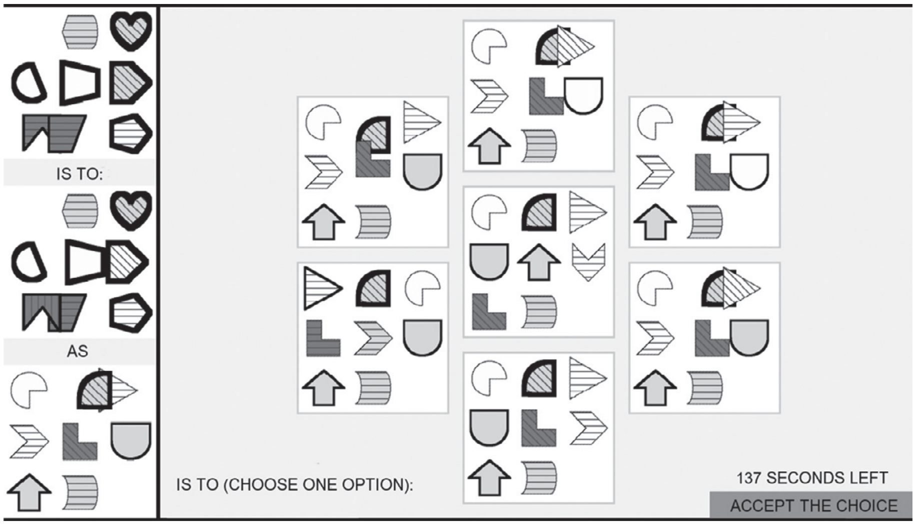

Figure 1. The high perceptual complexity (eight objects per analog) FAN test item, which includes the unary feature transformation between the top-left (A) and middle-left (B) patterns for the pair which becomes adjacent (a change in the saturation of one shape) as well as for the pair in which one shape covers another shape (a change in stripes). The task is to induce the types of location transformations used in A/B, and apply it to the bottom-left pattern C and the existing response options. See main text for details.

The right panel contained seven patterns arranged in a circular manner, which formed potential response options. The task was to choose with the mouse pointer the one and only option which correctly completed the analogy (stood for "?"), and then to accept that choice by clicking the "accept" button on the screen. Participants were given maximum 4 min. for each item.

Each pattern consisted of either 5 or 8 shapes (the low/high perceptual complexity condition, respectively), drawn randomly from a set of 16 predefined asymmetrical shapes, filling a virtual matrix of size of either 3 × 2 or 3 × 3, respectively. Each shape was characterized by four visual features: (a) orientation (0, 90, or 180° rotation in relation to the original shape), (b) saturation (white, light gray, dark gray), (c) the width of the border (thin, medium, thick), and (d) stripes inside (vertical, horizontal, oblique, or no stripes). Pattern B was always based on pattern A, with changes to the location of two pairs of shapes (so-called key pairs), according to one location transformation rule for each key pair, and one to four feature transformations in such a pair (the relational complexity condition: unitary, binary, ternary, and quaternary feature transformations could be used). There was a constraint that in one shape a maximum of three features could be altered. Either one or four shapes (depending on the perceptual complexity condition) were always irrelevant for the analogy (were outside the key pairs). There were four possible location transformation rules: either (a) the neighboring or (b) extreme shapes from a row/column could switch their locations, (c) two adjacent shapes could be separated or two nonadjacent neighboring shapes could be made adjacent, and (d) a shape partially covering a neighboring shape could become covered by that shape or vice versa (see Figure 1 for a sample of the FAN test item).

Pattern C included the same number of shapes as did A/B patterns, which either could be different or the same (on random) with regard to A/B. The sole correct response option was the result of applying to pattern C the same location transformations that changed pairs locations in A/B, and the same feature transformations for each pair as in A/B. One part of the task consisted of the correct detection of the key pairs in pattern C, that is, participants had to identify the shapes whose locations in C and in some response options altered according to the same rules as did the locations in A/B. The other part of the task included the choice of the one response option in which the same types of features of the key pairs changed as in A/B (it mattered which features changed, but not to what particular values). During the extensive instruction phase, as well as in the two training trials, participants were informed about the rules of the test.

A requirement to twice bind the same number of transformed features to each key pair was used in order to make the test be relatively difficult (as a primary interest was in the patterns of errors), as well as to allow for the possibility of creating more combinations of error options than would have been allowed by the use of only one key pair. It was assumed that each key pair could be processed in a separate step (i.e., the correct processing of the first pair narrowed the set of acceptable options, while the correct application of the second pair reduced that set to one option).

There were two incorrect response options for each of three investigated types of errors. The first type (corresponding to Vodegel Matzen et al's single omissions; henceforth called *almost-complete solutions*) consisted of options which matched the correct option with the sole exception that they missed one transformed feature for one random key pair. In the second type of error (*incomplete solutions*), the key pairs matched, but only the locations of shapes in one random pair were transformed properly and all the features were the same as in pattern C. So, these options were very similar perceptually to pattern C. In the last type of error, *non-relational solutions* (in a way corresponding to Raven et al.'s arbitrary-line-of-reasoning errors), the locations of the key pairs from pattern C changed according to location transformation rules different than rules used in A/B (either an A/B rule could be applied to wrong shapes or a different rule could be used). In such a kind of solution, four features also altered, but again they were not the proper features. So, these options were completely mismatched relationally, and different perceptually.

It was assumed that the choice of an almost-complete solution indicated that a participant followed the correct line of reasoning, but at its end failed to bind all features to respective shapes (i.e., missed one such feature). In the case of incomplete solutions, the most likely interpretation suggests that participants, when choosing such options, did not perform reasoning, but matched the response option on the basis of some simple heuristics, most probably the perceptual one ("do the patterns look similar or not?"). The non-relational/non-perceptual solutions acted as control options – their frequent choices might suggest that participants use random guessing to select options.

The design of the FAN test included eight conditions (with four test items per condition), as a result of the random manipulation of two factors: two levels of perceptual complexity, and four levels of relational complexity. Varying both perceptual and relational complexity of items was primarily meant to provide sufficient variation in difficulty of test items. I also expected that the choices of relationally close options (i.e., almost-complete solutions) would be sensitive to relational complexity, whereas the rate of selecting perceptually close options (i.e., incomplete solutions) would depend on perceptual complexity. Observing such facts will support the above interpretation of processing generating distinct types of errors. The dependent variables were proportions of *almost-complete-solution*, *incomplete-solution*, and *non-relational-solution* errors.

Benchmark Measures of Fluid Reasoning

Two more "traditional" paper-and-pencil tests of reasoning, the above-mentioned Raven's Advanced Progressive Matrices (Raven et al., 1983), and the figural analogy test (see Chuderski & Nęcka, 2012), were also applied. The latter test was especially designed to match the mean score and its standard deviation observed in Raven. The test includes 36 figural analogies in the form "A is to B as C is to X," where A, B, and C are types of relatively simple patterns of figures, A is related to B according to between two and five latent rules (e.g., symmetry, rotation, change in size, color, thickness, number of objects, etc.), and X is an empty space. The task is to choose one figure from a choice of four which relates to figure C, as B relates to A. Responses for both tests were recorded on an answer form. The total number of correctly answered items, in 60 and 45 min, was the score on Raven and the analogy test, respectively.

WM, STM, and Executive-Attention Tasks

Adapted versions of three complex span tasks: the operation span, reading span, and symmetry span tasks (Conway et al., 2005) were applied. Each task required participants to memorize a sequence of three to seven (i.e., set size) stimuli. Each stimulus was presented for 1.2 s apiece, out of nine possible stimuli for that task. Each stimulus was followed by a simple decision task, presented until a response was given, but for the maximum of 9 s. After two two-stimuli training trials, three trials for each set size (in increasing order) were presented in each complex span task. The operation span task analog required the memorization of letters, while deciding with a mouse button if an intermittent simple arithmetical equation (e.g., "$2 \times 3 - 1 = 5$?") is correct or not. The modified reading span task consisted of memorizing digits, while checking if letter strings (e.g., "EWZTE," "KTANY") begin and end with the same letter. The spatial span task involved memorizing locations of a red square in the 3×3 matrix, while deciding which of two presented bars is larger. The response procedure in each task consisted of a presentation of as many 3×3 matrices as was a particular set size, in the center of the computer screen, from left to right. Each matrix contained the same set of all nine possible stimuli for a given task. A participant was required to point with the mouse at those stimuli that had been presented in a sequence, in the correct order. Only a choice that matched both the identity and ordinal position of a given stimulus was taken as the correct answer. The dependent variable for each complex span task was the proportion of correctly pointed stimuli to all stimuli in the task. Complex spans are believed to reflect both the storage-capacity and executive-control aspects of WM (Unsworth & Engle, 2007).

In order to tap primarily the storage-capacity aspect of WM, a modified change detection paradigm (Luck & Vogel, 1997) was used, in three tasks: the letter, number, and color versions. Each of the 60 trials of the task (plus two training trials) consisted of a virtual, four-by-four array filled with a few stimuli (i.e., only some cells in the array were filled). The stimuli were 10 Greek symbols (e.g., α, β, χ, and so on), digits 0 to 9, or squares in 10 sufficiently distinctive colors, respectively. Each stimulus was approximately 2×2 cm in size. The number of stimuli within the array could be five, seven, or nine items. The array was presented for a period equal to the number of its items multiplied by 200 ms, and then followed by a black square mask of the same size as the array, presented for 1.2 s. In a random 50% of trials, the second array was identical to the first one, while in the remaining trials both arrays differed by exactly one item at one position, which was always a new item (not an item from another position). If the arrays differed, then the new item was highlighted by a square red border. If they were identical, a random item was highlighted. The task was to press one of two response keys depending on whether the highlighted item differed or not in two arrays. The second array was shown until a response was given or 4 s elapsed. The trials were self-paced. The score on this task is the estimated sheer capacity of the STM buffer (the k value; Cowan, 2001), calculated as the difference between the proportions of correct responses for arrays with one item changed and incorrect responses for unchanged arrays, multiplied by the set size. The total score was the mean k in the task.

Each antisaccade task, meant to capture primarily the executive-control aspect of WM, consisted of 5 training and 40 test trials. In order to increase the load on attention control, the tasks were slightly modified in comparison to the most commonly applied version (e.g., Unsworth, Schrock, & Engle, 2004): there were three locations (instead of one) to which a participant's eyes could be directed. Each test trial consisted of four events. First, a cue presented for 1.5 s informed that a target would be presented in the top, middle, or bottom of the side opposite to a flashing square (e.g., "Look at the bottom corner opposite to the flashing square," in Polish). Next, a fixation point was presented at the center of the screen for 1–2 s. Then, a rapidly flashing black square (3 cm in size) was shown in the middle of the left or right side of the screen, about 16 cm from the fixation point, for 0.15 s. Finally, depending on the task, a small dark gray arrow pointing left, down, or right (a spatial version), a digit "1," "2," or "3" (a number version), or a string "left," "down," "right" (a letter version), was presented in the location opposite to the square for only 0.2 s and was then replaced by a mask. The task was to look away from the flashing square, to detect the direction of the arrow or the identity of the digit/string, and to press the key associated with the stimulus. The trials were self-paced. The dependent variable in each task was mean accuracy.

Procedure

The study was a part of a larger project, which consisted of 17 computerized WM tasks, related also to other research goals (for details see Chuderski, 2014). Apart from three WM, three STM, and three antisaccade tasks that were described above, it also included three *n*-back tasks, two Stroop tasks, the stop signal task, and two relation monitoring tasks. Because the three former types of tasks yielded very moderate correlations with Gf, whereas the latter task was aimed to specifically tap the binding and relation integration processes, these tasks are not reported here. The whole WM session lasted about 4 hr.

In another session, participants completed, also unreported here, a paper-and-pencil relation discovery test (see Chuderski, 2013), and then the three fluid reasoning tests, applied in the following order: FAN, Raven, and paper-and-pencil analogies. The order of sessions was random. They were separated by a 1-hr lunch break.

Results

Table 1 presents the summary of measures. The effects yielded by the two complexity manipulations in the FAN

Table 1. Descriptive statistics and reliabilities for measures used in the study

Task	*M*	*SD*	Range	Skew	Kurtosis	Reliability
Total error rate in FAN	0.54	0.25	0.00 to 1.00	−0.34	−0.94	.90
Almost-complete solution rate	0.35	0.15	0.00 to 0.69	−0.19	−0.44	n/a
Incomplete-solution rate	0.12	0.15	0.00 to 0.84	1.94	4.77	n/a
Non-relational solution rate	0.07	0.08	0.00 to 0.41	1.41	1.91	n/a
Almost-complete solution prop.	0.72	0.22	0.09 to 1.00	−0.55	−0.42	n/a
Incomplete-solution prop.	0.16	0.17	0.00 to 0.84	1.25	1.57	n/a
Raven APM	22.38	6.34	3 to 35	−0.58	0.23	.87
Paper-and-pencil analogies	22.52	6.28	6 to 35	−0.24	−0.64	.83
Fluid reasoning factor	0.00	0.91	−2.68 to 1.99	−0.34	−0.25	n/a
Raven APM – AC group	24.91	5.52	4 to 35	−0.78	1.15	.85
Paper-and-pencil analogies – AC	24.58	5.81	7 to 35	−0.48	−0.36	.81
Raven APM – IS group	19.81	6.09	3 to 32	−0.51	−0.01	.85
Paper-and-pencil analogies – IS	20.44	6.06	6 to 34	−0.03	−0.53	.81
Color STM	2.85	1.38	−0.87 to 6.00	−0.30	−0.37	.67
Letter STM	2.44	1.42	−1.33 to 6.40	−0.09	−0.27	.68
Number STM	4.45	1.47	−0.77 to 7.00	−1.09	1.09	.80
Spatial antisaccades	0.60	0.24	0.05 to 1.00	−0.31	−1.05	.92
Letter antisaccades	0.69	0.25	0.00 to 1.00	−0.74	−0.60	.94
Number antisaccades	0.48	0.22	0.00 to 1.00	0.04	−0.74	.90
Spatial complex span	0.52	0.18	0.05 to 0.97	0.16	−0.10	.84
Letter complex span	0.70	0.18	0.05 to 0.99	−0.72	0.11	.87
Number complex span	0.78	0.15	0.09 to 1.00	−1.18	2.03	.84

Notes. $N = 347$ for the error rates in FAN, and for Raven APM, Analogies, Fluid reasoning factor, and all WM tasks. $N = 344$ for both proportions (as three people committed no errors, no proportions could be calculated for them). $N = 175$ for the AC group. $N = 172$ for the IS group. Reliability = Cronbach's alpha. n/a = nonapplicable. Prop. = the proportion of errors in all errors. The AC/IS group stands for the group scoring below/above median in the proportion of incomplete-solution errors. STM = short-term memory. APM = Advanced Progressive Matrices.

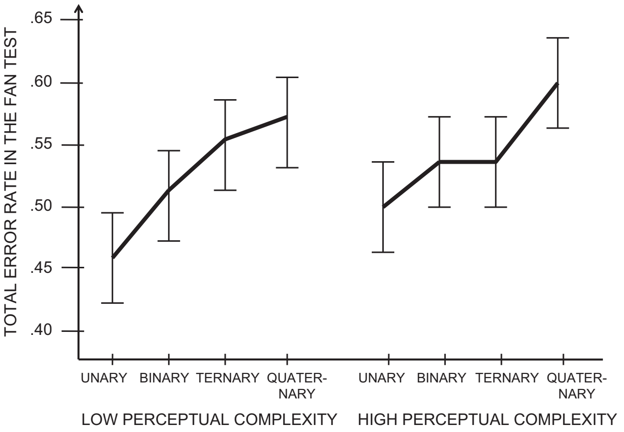

Figure 2. Total error rate in the FAN test as a function of relational complexity (unary to quaternary), and perceptual complexity (five- or eight-element items). Bars represent 95% confidence intervals.

test were analyzed first. Figure 2 summarizes the error rate in all conditions of this test. A 2 × 4 ANOVA indicated that there were more errors under the high perceptual complexity ($M_8 = .56$) than under the low one ($M_5 = .52$), $F(1, 242) = 18.74$, $p < .001$, $\eta^2 = .054$, as well as there were more errors with the increasing level of relational complexity ($M_1 = .48$, $M_2 = .54$, $M_3 = .55$, and $M_4 = .58$), $F(3, 1{,}038) = 25.16$, $p < .001$, $\eta^2 = .067$. The error rate varied from $M = .46$ to $M = .60$ between the test conditions. Thus, the influence of complexity factors on accuracy was significant, but moderate.

Participants made significantly more almost-complete-solution than incomplete-solution errors, $t(345) = 21.05$, $p < .001$, Cohen's $d = 28.87$, as well as made more of the latter type errors than non-relational-solution errors, $t(345) = 5.16$, $p < .001$, $d = 6.95$ (so, as these solutions were committed very infrequently, and they have no theoretical interpretation, they were not analyzed further). Contrary to frequent observations (e.g., Lynn & Irwing, 2004) of sex differences in the Raven test (i.e., males outperform females), no such differences in the total FAN score nor in the particular error types were found (the same observation pertained also to the Raven and paper-and-pencil analogy scores).

In order to test the patterns of errors, the proportions of a given type of errors in the total number of errors were calculated. The proportion of almost-complete solutions in all errors did not change significantly with increasing perceptual complexity ($M_5 = .72$ vs. $M_8 = .71$), $t(330) = 0.60$, whereas the proportion of incomplete solutions did ($M_5 = .15$ vs. $M_8 = .18$), $t(229) = 2.59$, $p = .010$, $d = 0.15$. The relational complexity factor significantly affected almost-complete solutions ($M_1 = .73$, $M_2 = .71$, $M_3 = .66$, $M_4 = .65$), $F(3, 870) = 10.01$, $p < .001$, $\eta^2 = .033$, whereas, when taking into account the large sample examined, the effect of relational complexity on incomplete solutions ($M_1 = .17$, $M_2 = .18$, $M_3 = .20$, $M_4 = .20$) was in fact negligible, $F(3, 870) = 2.84$, $p = .036$, $\eta^2 = .001$ (note that only participants who committed at least one error in each level of a given factor were analyzed above).

The most important analyses regarded correlations between reasoning ability factor, calculated as the mean of *z* scores in the two benchmark reasoning tests, and the proportions of errors. The total error rate correlation with ability factor was $r = -.524$, $p < .001$. The proportion of almost-complete solutions in all errors significantly increased with increasing ability, $r = .541$, $p < .001$, whereas the proportion of incomplete solutions accordingly decreased with increasing ability, $r = -.482$, $p < .001$. Both these relationships between ability and the error proportions are depicted in Figure 3. The absolute strength of correlation between the ability factor and the proportion of almost-complete solutions in all errors was significantly

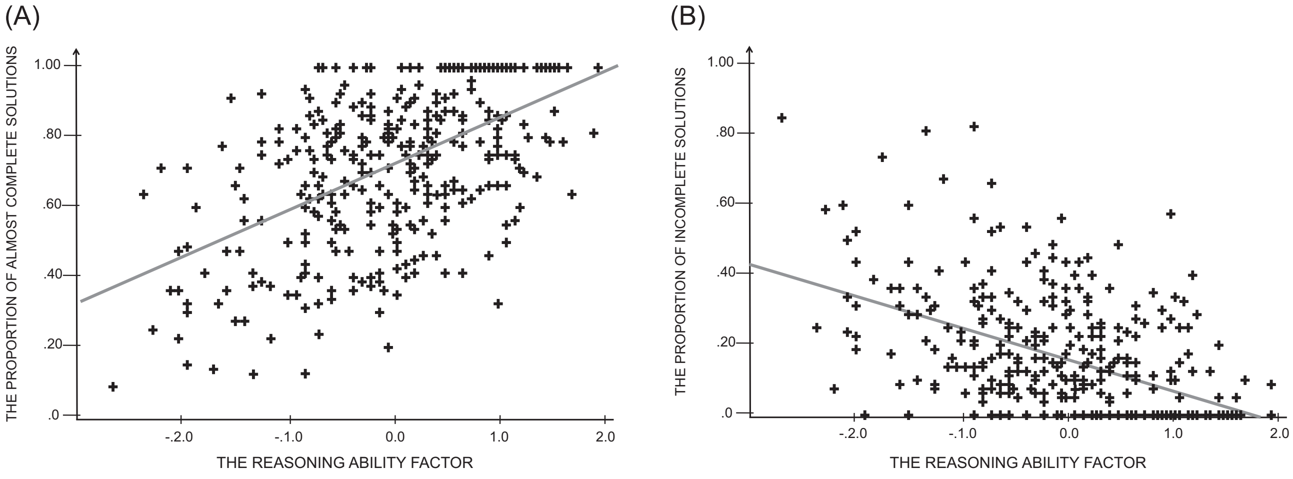

Figure 3. The scatterplot of the proportion of the almost-complete solutions (A) and the proportion of incomplete solutions (B) in all errors in the FAN test, as a function of the reasoning ability factor. Solid lines represent regression lines.

Table 2. The matrix of correlations among scores used in the confirmatory factor analysis

Task	1	2	3	4	5	6	7	8	9	10
1. Color STM										
2. Letter STM	.401									
3. Number STM	.397	.471								
4. Spatial antisaccades	.391	.403	.410							
5. Letter antisaccades	.410	.406	.426	.819						
6. Number antisaccades	.381	.355	.366	.782	.686					
7. Spatial complex span	.401	.340	.374	.491	.454	.387				
8. Letter complex span	.322	.384	.352	.438	.417	.353	.532			
9. Number complex span	.281	.318	.404	.400	.463	.317	.465	.688		
10. Raven APM	.425	.338	.370	.435	.408	.403	.478	.429	.381	
11. Analogies	.347	.248	.255	.398	.384	.380	.433	.446	.365	.656
Raven APM – AC group	.176	*.124*	.342	.322	.343	.359	.363	.333	.219	
Analogies – AC group	.151	*.095*	.211	.268	.229	.274	.339	.297	.200	.604
Raven APM – IS group	.516	.424	.314	.389	.415	.415	.422	.327	.348	
Analogies – IS group	.405	.287	.204	.388	.404	.374	.379	.429	.363	.606

Notes. $N = 347$ for all tasks. All correlations are significant at the $p < .05$ level, except for two values presented in italics. Four bottom rows present data for the groups scoring below (AC) and above (IS) median in the proportion of incomplete-solution errors. STM = short-term memory. APM = Advanced Progressive Matrices.

larger, $t(345) = 7.49$, $p < .001$, than the strength of a respective correlation between the ability factor and the bare number of almost-complete solutions, $r = -.174$, $p = .001$ (in the case of the number of incomplete solutions r equaled $-.481$). The proportion of errors was also an excellent predictor of the total number of errors in the FAN test, as the proportion of incomplete solutions increased with increasing number of errors, $r = .719$, $p < .001$ (of course the proportion of almost-complete solutions decreased accordingly, $r = -.731$, $p < .001$).

Interestingly, the segmented linear regression analysis (for a description of the use of this method in intelligence research see Jauk, Benedek, Dunst, & Neubauer, 2013), which estimated regression coefficients separately for participants falling either below or above median in reasoning factor (i.e., median Gf was fixed as a breakpoint parameter in a regression model), indicated that in the below-median participants an increasing reasoning factor predicted the number of almost-complete solutions positively, $\beta = .300$, $p = .004$, whereas in the above-median participants the former factor predicted the latter number of errors negatively, $\beta = -.786$, $p < .001$. No such relationship existed for the number of incomplete solutions, which were negatively related to the reasoning factor in both the below-median, $\beta = -.669$, $p < .001$, and above-median participants, $\beta = -.362$, $p = .001$, though the former relationship was significantly stronger, $t(345) = 2.99$, $p = .001$. No difference between the below- and above-median participants was found using the segmented regression analysis for the proportion of each type of errors in all errors.

The last analysis compared, with the use of confirmatory factor analysis (CFA), the strengths of links between the reasoning ability factor and the complex span, STM, and antisaccade factors, loaded by each type of WM task used (note that assuming a lesser number of factors – one or two – yielded very poor fit of respective models), separately for half participants who committed primarily the almost-complete solutions (i.e., the proportion of such errors in all errors was no less than $M = .89$) versus the other half of participants who committed a substantial proportion of the incomplete solutions (i.e., larger than $M = .11$). Due to the above-mentioned strong correlation between this latter proportion and the total error rate in the FAN test, the half of participants who committed more incomplete solutions included primarily people who scored on FAN below median. A group CFA model was calculated (zero-order correlations underlying this model are presented in Table 2), in which the correlations between WM factors and the Gf factor were calculated separately for each group, as it was expected that WM tasks would constitute stronger predictors of Gf in the "incomplete-solutions group" (IS group) than in the "almost-complete-solutions group" (AC group). However, at the same time the model assumed in both groups the equal loadings of each factor on respective manifest measures (i.e., scores in WM tasks/Gf tests), as there is no reason to expect that particular tasks reflect the underlying constructs differently in each group. Also, the particular links between WM factors were assumed equal in both groups.

The resulting model is presented in Figure 4. The fit of the model was good, $N = 347$, $\chi^2(90) = 170.07$, $\chi^2/df = 1.89$ (the widely accepted criterion value is $\chi^2/df < 2.0$; for this and subsequent criteria see Hu & Bentler, 1999; Kline, 1998), adjusted population gamma index = .960 (criterion value > .92), RMSEA = .071 [.054–.087] (criterion value < .08), SRMR = .069 (criterion value < .08). The most important result indicated that each WM factor significantly correlated with the Gf factor (all ps < .001), but for each WM factor the correlation coefficient was significantly larger in the IS group than in the AC group: $t(345) = 2.11$, $p = .018$, $t(345) = 4.02$, $p < .001$, and $t(345) = 2.60$, $p = .004$, for the complex span, STM, and antisaccade factors, respectively.

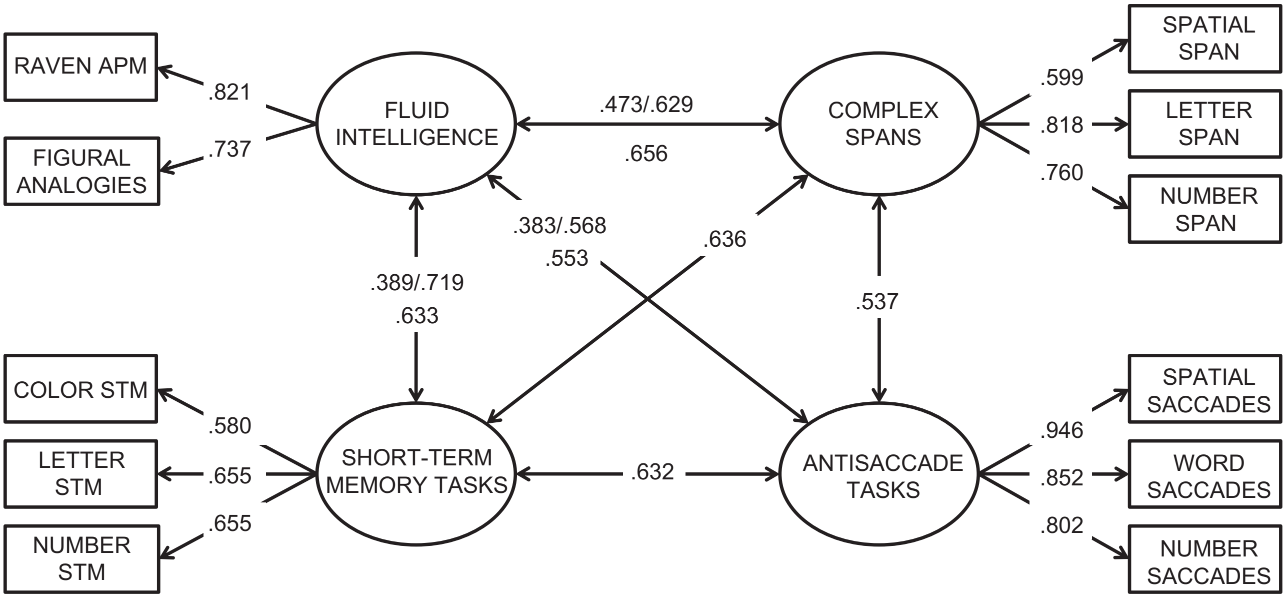

Figure 4. The confirmatory factor analysis, including latent variables (ovals) representing variance shared by two reasoning tests and triples of the complex span, short-term memory (STM), and antisaccade tasks, respectively. Boxes represent particular tasks. Values between ovals and boxes represent relevant standardized factor loadings (all *p*s < .001). Values between ovals represent relevant correlation coefficients among latent variables for the whole sample (all *p*s < .001). Pairs of values separated by "/" reflect correlation coefficients among Gf and WM variables for the groups scoring below/above median in the proportion of incomplete-solution errors (all *p*s < .001).

General Discussion

To sum up the presented study, its most important result indicates that the pattern of errors that people committed, when dealing with the specially designed, computerized test of figural analogy, strongly predicted their fluid intelligence level measured by the two standard, paper-and-pencil tests. The patterns of errors were a no worse Gf predictor (|*r*| = .541) than just the total error rates (|*r*| = .524). Regression and correlational analyses suggested that the average reasoners differed from the best reasoners primarily by a larger number of almost-complete solution options, what indicated that both the former and the latter followed the correct line of reasoning, attempting at finding and applying the relations included in an analogical problem, but the average reasoners more often missed one relational element from the complete solution. In contrast, poor reasoners committed even less almost-complete-solution errors than did average reasoners, but at the same time the former accepted a much larger number of incomplete solutions than the latter. This fact suggests that poor reasoners often did not engage in a reasoning process (if they did, they would easily notice that a few relations lacked in an incomplete solution), but relied on some simpler heuristic.

The analysis of variance related to experimental manipulations applied to the FAN test items suggests that, in cases when poor reasoners were not able to rely on their WM, such a heuristic consisted of finding the response option most similar perceptually to the target option. First, incomplete solutions, most frequently chosen by the poorest reasoners, were the only options sensitive to increased perceptual complexity. As eight-element incomplete solutions shared more identical elements with targets, this fact might strengthen the poorest reasoners' relying on the perceptual heuristic, in comparison to dealing with five-element items. No analogous sensitivity of incomplete solutions pertained to relational complexity (whereas the latter factor did affect almost-complete solutions that probably were more closely linked to relational reasoning). Second, non-relational solutions, which were not as much perceptually similar to targets as were incomplete solutions, were chosen relatively infrequently, even by poor reasoners. So, the random guessing heuristic can most likely be rejected as a primary basis of incorrect decisions on the FAN test.

The fact that poorly reasoning participants differed from the best reasoners in a qualitatively different way than did medium-ability people is consistent with the observation made by Forbes (1964), but quite discrepant to null findings of Vodegel Matzen et al. (1994), showing the similar patterns of errors for the low-, medium-, and high-ability children. The index of the pattern of errors expressed as the proportion of almost-complete solutions was the numerically best predictor of the scores on benchmark tests of fluid reasoning, indicating that with increasing fluid intelligence people not only committed less reasoning errors, but also were closer to the correct solution, even if they eventually committed an error.

This result empirically supports the model of Carpenter et al. (1990), who predicted that worse reasoners often lose track in their WM of correct rules that need to be applied, so these reasoners may often choose options that do not follow those rules. At the same time, better reasoners effectively manage the necessary rules in WM, and when they make errors, they make them for reasons other than just losing the representation of a problem from WM. For instance, according to Carpenter et al., those reasons may include insufficient abstraction or failed control over processing. Moreover, Carpenter et al.'s conclusion was strongly supported by the CFA that tested the WM-Gf correlations separately for halves of participants who made relatively less versus more incomplete-solution errors, that is, who where on average closer versus farther from the correct solutions. The performance on two benchmark tests of fluid reasoning was much less dependent on WM resources in the case of the former half (i.e., WMC measured by complex spans explained 22.3% of Gf in that group) than in the case of the latter half (i.e., 39.5% Gf variance was explained by WMC).

It also seems that the effect of group on the WM-Gf link was more substantial for measures reflecting primarily storage capacity (i.e., STM factor explained as much as 51.6% of Gf variance in the IS group) than for tasks meant to tap executive control (32.3% Gf variance explained by the antisaccade tasks in that group). This observation is also consistent with Carpenter et al. model.

However, the present study is less conclusive with regard to existing alternative theories of mechanisms underlying Gf that have been adopted in correlational studies on intelligence. The presented results can be potentially explained in terms of both the executive-control and storage-capacity accounts of fluid intelligence. A possible interpretation (e.g., Robin & Holyoak, 1995) of the high proportion of almost-complete solutions, accounting for the differences in executive control, might tell that such errors are made when participants cannot inhibit responses based on partial solutions that match a certain number of correct features ("I found a match, so I am not checking more precisely"). The alternative explanation in terms of insufficient storage capacity might predict that, in order to be correctly integrated with the appropriate objects, all feature transformations and a pair of objects must be actively maintained in parallel in WM (Carpenter et al., 1990; Oberauer et al., 2007). Highly capacious participants may lack just one memory slot for just one transformation, and so if they make errors, they usually choose options missing only one feature. WM of less capacious people may not fit more than two or three slots (Cowan et al., 2006), so actively maintaining two objects constituting a correct relation plus up to four transformations may be impossible for such participants, and so they chose options missing more features.

Unfortunately, the present results do not allow for deciding between these two alternative accounts, as measures of both STM and executive control (i.e., the antisaccade tasks) provided comparable strengths of correlation with Gf in the full sample (and similar to that of complex spans). Only the data from poorly reasoning participants supports the STM explanation relatively more than it supports the executive-control account. Future studies on error patterns in fluid intelligence tests, linking these patterns to capacity and control aspects of WM, are required. Such studies should include more discriminative methods regarding the identification of possible mechanisms of error commitment (e.g., include error options which are chosen due to loading either working memory capacity or executive control). It is also possible that both executive control and storage capacity underlie fluid intelligence (see Chuderski & Necka, 2012; Conway, Getz, Macnamara, & Engel de Abreu, 2011), for instance in interaction, or depending on the adopted strategy of dealing with the reasoning test (e.g., the correct-solution construction strategy may be more dependent on storage capacity, whereas the response-option elimination strategy may be more prone to distraction, and rely on executive control to a larger extent).

The interpretation of the present results in terms of two different modes of reasoning: one more simplistic (perceptually-based), and the other more elaborate relationally, is also consistent with the dual-process (DP) account of reasoning. The DP account assumes that the implicit, heuristic, and genetically determined associative mechanisms capture thinking unless deliberately overridden by an explicit, analytical, and properly trained mechanism which enables mental simulations and hypothetical reasoning (e.g., Evans & Over, 1996; Sloman, 1996; Stanovich & West, 2000). Of course, the DP account provides one possible explanation for the presented results, and they can also be reasonably (and more parsimoniously) explained by the single-process theories of reasoning, like the mental model theory (MM; Johnson-Laird, 1999, 2006). The MM theory would predict that people who get captured by incomplete solutions in the FAN test construct too simple or too few (initial) mental models, which do not allow them to represent counterexamples that would refute the distractor as a proper solution. In contrast, participants who construct more elaborate models and thus represent counterexamples may easily avoid incomplete solutions, but their either insufficient executive control or storage capacity may not suffice for the eventual construction of the correct solution, and they may end up with an almost-complete solution.

It is also possible that some participants solely used the MM machinery for reasoning, with varying degrees of success, while others relied on a mixture of heuristics and MM. In fact, Oberauer (2006) contrasted both such models (i.e., MM vs. MM plus heuristics) with data on reasoning with conditionals, and both models displayed a similar fit. The MM model assessed that 26% participants just guessed an answer, 27% made inferences only from the initial model, while the remaining 47% of participants performed the hypothetical thinking. The DP model assumed that 6% people guessed an answer, 51% used the heuristic process, while the remaining 43% constructed elaborate mental models. So, both accounts predict that almost 50% of participants will utilize a reasoning process, while the other 50%

(or more) of participants will rely on a simpler heuristic (most likely because of the lack of necessary WMC). These predictions are in general agreement with the present data.

To conclude, the current paper contributes to the research on fluid intelligence by providing two original observations on how young adults differ in why they fail on intelligence tests. First, in line with the seminal model of reasoning proposed by Carpenter et al. (1990), for the first time in the literature it has been showed with a sufficient sample that good reasoners followed the rules governing analogies to be solved, but sometimes omitted small fragments of the correct solution. Poor reasoners much more often ignored or lost track of the required rules, and chose options that were far from the correct solution. Second, the study has yielded novel data showing that in the group of poor reasoners there was a stronger relationship between available WM resources and scores on the benchmark intelligence tests than was in the good reasoners. These two results support the existing influential model of performance in intelligence tests (Carpenter et al., 1990), and provide us with important insights into processes and mechanisms possibly responsible for reasoning ability that may differentiate intelligent people from less intelligent ones.

In general, the results presented above show that a lot can be learned and understood about the nature of fluid intelligence from the analyses of properly designed errors in intelligence tests, combined with correlational analyses, which now dominate the field, and so far rarely undertaken computational modeling. The work suggests that correlational studies on the cognitive mechanisms of intelligence might benefit from being more often supplemented with detailed analyses of reasoning processes carried out during intelligence tests, bringing differential psychology closer to cognitive psychology, as advocated by Cronbach (1957).

Acknowledgments

This study was supported by Grant No. N106 417140 sponsored to A. Chuderski by the National Science Centre of Poland. Thanks to Krzysztof Cipora, Dominika Czajak, and Jolanta Wójcik for their help in conducting the study and to Mike Timberlake for correcting the text.

References

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's Advanced Matrices Test. *Intelligence, 30*, 485–503.

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

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General, 140*, 674–692.

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.

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence, 41*, 244–262.

Chuderski, A. (2014). Relational integration task explains fluid reasoning above and beyond other working memory tasks. *Memory & Cognition, 42*, 448–463.

Chuderski, A., & Necka, E. (2012). The contribution of working memory to fluid intelligence: Capacity, control, or both? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*, 1689–1710.

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence while executive control does not. *Intelligence, 40*, 278–295.

Colom, R., Abad, F. J., Quiroga, M. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs but why? *Intelligence, 36*, 584–606.

Conway, A. R. A., Getz, S. J., Macnamara, B., & Engel de Abreu, P. M. J. (2011). Working memory and fluid intelligence: A multi-mechanism view. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 394–418). Cambridge, UK: Cambridge University Press.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–114.

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition, 34*, 1754–1768.

Cronbach, L. J. (1957). Two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Evans, T. G. (1968). A program for the solution of a class of geometric analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing* (pp. 271–353). Cambridge, MA: MIT Press.

Forbes, A. R. (1964). An item analysis of the advanced matrices. *The British Journal of Educational Psychology, 34*, 1–14.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences, 11*, 236–241.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence, 40*, 427–438.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence, 41*, 212–221.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology, 50*, 109–135.

Johnson-Laird, P. N. (2006). *How we reason?* Oxford, UK: Oxford University Press.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). Oxford: Oxford University Press.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 66–71.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research, 22–23*, 47–66.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 2761–2766). Austin, TX: Cognitive Science Society.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281.

Lynn, R. L., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence, 31*, 481–498.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology, 53*, 238–283.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). Oxford, UK: Oxford University Press.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and vocabulary scales (Section 4: Advanced Progressive Matrices)*. London, UK: H. K. Lewis.

Robin, N., & Holyoak, K. J. (1995). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 987–997). Cambridge, MA: MIT Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645–726.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132.

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1302–1321.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influence on intelligence. *Intelligence, 34*, 261–272.

Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences, 16*, 433–445.

Date of acceptance: November 10, 2014
Published online: July 1, 2015

Adam Chuderski

Cognitive Science Department
Jagiellonian University
Grodzka 52
31-044 Krakow
Poland
E-mail adam.chuderski@gmail.com