Why People Err on Multiple-choice Analogical Reasoning Tests

Adam Chuderski
Jagiellonian University in Krakow
Adam.Chuderski(at)uj.edu.pl

Bartłomiej Kroczek
Jagiellonian University in Krakow
Bartek.Kroczek(at)gmail.com

Abstract

A widespread tool in analogy research consists of multiple-choice tests that require identifying a relation between two situations and mapping it to another two situations, to find the correct response option. A key source of difficulty during such tests is attributed to the complexity of mapping. However, most people do not construct mappings purely in the mind, but also compare the emerging mapping with the existing response options, so their features may affect the reasoning process. This study examined the impact of relational match of error options with respect to the correct option (the proportion of correct elements present in a given error option) on the option selection. Results indicate that option selection depends almost linearly on the relational match. Moreover, the higher working memory capacity of the participants, the more relationally matching errors they select. The study suggests careful design of error options in multiple-choice reasoning tests, because the pattern of these options can affect the solution process.

Keywords: analogy, reasoning tests, errors, working memory

Introduction

Analogy is a crucial general-domain cognitive mechanism underlying the transfer of information from the known (*source*) to the unknown situation (*target*), operating at various levels of the cognitive system, from perception to reasoning and problem solving. The key process in analogy making, called *mapping*, consists of finding the systematic correspondence between relations that validly describe the source and the target. According to an influential theory of mapping (Gentner, 1983), people tend to look for the most comprehensive (explanatory) mapping available that can describe as many objects, their attributes, and their roles in relations as possible but that, at the same time, is maximally univocal and productive. When the correct mapping of relations has been made, unknown elements in the target can be filled in using the known elements from the corresponding places in the source, by means of *analogical transfer* (see Holyoak, 2012). Analogical mapping and transfer have been studied intensively for the last fifty years.

Analogy making performance is typically examined using various analogical reasoning tests. For instance, in the scene analogy task (e.g., Gentner & Toupin, 1986; Richland, Morrison, & Holyoak, 2006), a participant is shown the source scene in which a straightforward relation is depicted (e.g., a cat chasing a mouse). The task is to identify an object in the target scene (e.g., a boy chasing a girl) that corresponds relationally to a given object in the source scene (e.g., the boy = the cat). Such tasks become more difficult when the relation includes more arguments (relational roles), as in mapping a dog chasing a cat chasing a mouse onto a women chasing a boy chasing a girl. Error rates also rise when distraction occurs, as in mapping a dog chasing a cat chasing a mouse onto a boy chasing a dog chasing a cat (the two dogs cannot be mapped together, because they play different roles).

Most of analogical reasoning tests, used in the analogy research, exploit the four-term format: a relation has to be identified between the terms A and B, and then applied to term C in order to establish term D that satisfies the relation. For example, for the relation *a part of*, a leg is to a body as a door is to a car. Such semantically-based analogies are relatively easy to solve for healthy adults, and thus are used primarily in research on children, elderly, and clinical samples (e.g., Krawczyk et al., 2008; Thibaut & French, 2016). To study analogical reasoning in healthy adults, geometric four-term analogies are typically applied, in which relations are defined formally (e.g., as geometric properties of shapes), and cannot be easily identified on a basis of familiarity or common knowledge (Bethel-Fox, Lohman, & Snow, 1984; Novick & Tversky, 1987). Because difficulty of such tests is considerable, constructing the solution (D) term from scratch is barely possible (but see Lovett, Tomai, Forbus, & Usher, 2009), thus commonly a number of alternative response options is presented to a participant, who has to select the one and only correct option out of that set.

Several studies (e.g. Hosenfeld, van der Maas, & van den Boom, 1997; Primi, 2002) suggested that difficulty of the multiple-choice test items is driven by their relational complexity. However, it was also noted that the items' relational complexity is commonly confounded with their perceptual complexity (more complex relations typically require a larger number of shapes and shape attributes; Primi, 2002). Therefore, a more univocal examination of the complexity factor is required (see below).

Moreover, other factors, beyond relational complexity, might affect the difficulty of geometric analogical reasoning. The objective of the present work was to examine a potential

role of one such factor: the characteristics of error response options that accompany the correct response option.

To this aim, a novel computerized geometric analogy test was developed, in which the correct response option (D) was surrounded by five erroneous options. They varied in the number of correct geometric transformations that were required to obtain D from C (i.e., those transformations which applied also to A and B), from having just one correct transformation, up to having all but one transformation. The proportion of the correct transformations defined *relational match* (RM) of options, with the correct option D by definition equaling RM = 100%, and the remaining options equaling RM below 100%. The main research question was: Would RM values of error options affect their selection rates?

Two data patterns were possible. First, if errors resulted primarily from not running analogical mapping, but from using some superficial solution strategy instead (e.g., following perceptual similarity; Kunda, McGreggor, & Goel, 2013), then the low-RM options should be selected more frequently than the high-RM options, as the former were more similar perceptually to option C, as well as their processing was simpler (e.g., required noticing a single transformation). Alternatively, if the participants involved in analogical mapping, and tried to identify and map as many correct transformations as possible (for them), the errors could primarily result from not completing the mapping (e.g., missing just a single transformation), and thus the high-RM options should be selected preferably over the low-RM ones.

However, even if this study failed to observe a general effect of RM, some individual differences in the tendency to select particular response options could still exist. In order to examine whether such differences could be meaningful, working memory capacity (WMC) of the participants was screened. WMC reflects individual effectiveness in active maintenance and manipulation of task-relevant information in the mind (Cowan, 2001). WMC strongly predicts proportion correct on analogical reasoning tests (see Holyoak, 2012), but it is less understood how WMC affects reasoning errors. Relatedly, the second research question was: In what way would WMC predict the mean RM values of errors that participants made?

Thirdly, to examine the impact of relational complexity on the errors committed, the number of objects transformed (and thus the number of transformations) was manipulated. However, crucially, the number of visual features present in options C and D was held constant, so any effects found could not be attributed to perceptual complexity, but solely to the geometric relation complexity.

Method

Participants

A total of 293 volunteers, recruited via ads on popular networking websites, attempted geometric analogies as well as three working memory tasks. After the initial inspection of the results, the data from 28 people (9%) were discarded either due to having accuracy at the floor level and a mean RT below 15 s (i.e., below a reasonable amount of time that was needed for a valid perception of the test item), indicating that such a person accepted options through rapid guessing, or due to choosing the option identical to C in the majority of trials, meaning that such a person most likely did not understand the task instruction. In both cases, most probably no analogical mapping was even attempted, so these data would yield noise if included in the analyses. The final sample consisted of 164 women and 101 men, aged 18–46 (M = 23.6 y, SD = 6.4 y), who were paid the equivalent of 20 euros in the local currency. All the participants were informed that the study was related to thinking and that their data would be anonymous. Participation could be ended at will at any moment. The study conformed to the ethical principles of the WMA's Declaration of Helsinki.

Geometric analogies

In each trial of a computerized test, an A:B::C:D geometric analogy was generated automatically. It consisted of terms A, B, and C, placed in one column on the left side of the screen (A over B, B over C), as well as the six response options (including the correct option D), displayed on the right side of the screen (for the general layout of the analogy, see Fig. 1). The two areas of the screen were separated by a black vertical line. Each term and response option was a gray square approx. 6 cm in size. It included four simple geometrical shapes, organized in the two by two layout. The instruction, placed on the screen, informed the participants that "A is to B like C is to: (choose one response option)" in the local language. A response option was selected by clicking on it with the computer mouse. A green rim surrounded the selected option. The selection could be changed before the deadline. A clock indicated the seconds remaining. The final selection was made by clicking on the button "Accept answer" (in the local language) before the deadline. If this button was not clicked on, a trial was treated as an error and was discarded from the latency analyses.

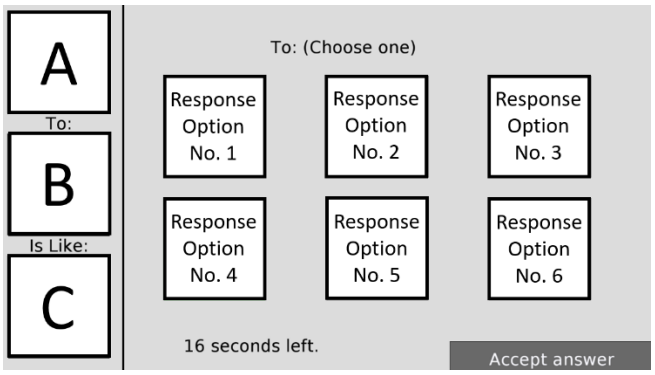

Figure.1: The general layout of the geometric analogy task

The following rules governed the A:B::C:D analogy. The terms A and B contained the same four shapes drawn from the pool of twenty maximally distinctive yet simple shapes. One to three visual features of as many as two, three, or four shapes (the *complexity* condition) were transformed in the term B relative to the respective features in the term A. Three feature transformations (hereafter, *transformations*) were allowed: a change in filling (white, light gray, or dark gray), a change in rim (thin, medium, or thick), and a change in rotation (by 90°, 180°, or 270°). As many as one, two, or three transformations could be applied to each transformed shape. The number of transformations for each shape was random. Each shape was transformed using the unique combination of transformations or the unique transformation. Specific levels of features did not matter for the task (e.g., if a shape's white filling changed to either light or dark gray); only the particular unique combinations of transformations for given shapes needed to be tracked by the participants.

In the example analogy presented in Fig. 2, the teardrop shape became rotated by 270°, its filling changed from dark gray to white, and its rim thickened (three transformations), the rounded triangle's filling changed from white to dark gray and its rim thickened (two transformations), and the trapezium's rim became thinner (a single transformation). To make the task more difficult (errors were the primary focus of the study), some of the shapes in B randomly changed locations relative to A (the teardrop and the trapezium).

The participants were instructed to detect which shapes had been transformed between A and B and to identify for each shape the unique combination of transformations to be applied to shapes in C, which comprised another four shapes drawn from the pool. The same four shapes were used in each response option (their location relative to C could change). The goal was to select the response option (*all-correct*) in which the same number of shapes in C as in A and B were changed according to the same transformation combinations.

In Fig. 1, the bottom-left option is correct, because three transformations were applied to one shape (the crescent), the filling and the rim were transformed for the second shape (the "L"), and the rim was transformed for the third shape (the horseshoe). The remaining five options lacked some of the correct transformations. One option (*all-1-invalid*) included the same combination of transformations as *all-correct*, except that one of them was different from that in the A-B pair (in the bottom-middle option, the horseshoe rotation was changed, instead of its rim). Another option (*all-but-1*) included all the correct transformations except that one transformation was missing (in the top-left option, the L shape's filling was not changed from white). In the bottom-right option (*just-2*), only two transformations were applied (only the crescent was rotated and its rim was changed). In the top-right option (*just-1*), only one transformation was applied (the rim of the L shape was altered). In the top-middle option (*no-transformation*), no shape was transformed at all, and the shapes looked the same as in C.

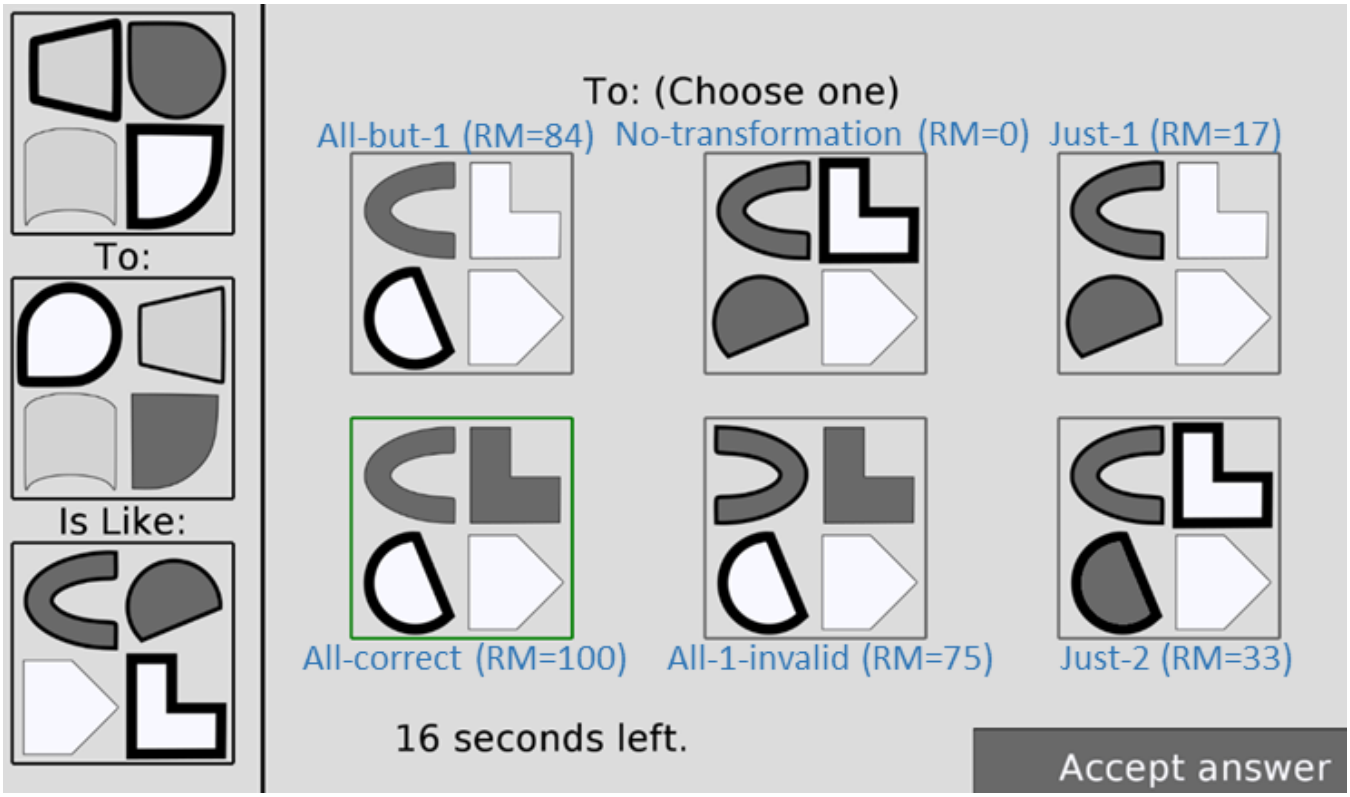

Figure 2: An example geometric analogy (legend in blue).

For each participant, six unique analogies per complexity condition were generated (in total, 18 analogies were applied per person). Before the main test started, a detailed written instruction explained the task, and an example screen visualized the transformations, highlighting the respective correct response. Research assistants provided additional explanations if necessary. Also, two training analogies of the lowest complexity were applied with a deadline of 150 s each. Accuracy feedback was given on these training analogies; there was no feedback on the main test.

The key dependent variable was the RM metric. RM of a given option equaled the proportion of the correct transformations present in that option, relative to all the correct transformations (for ease of presentation, the proportions were rounded to two decimal places and expressed as percentage values). For the correct option, by definition, its RM = 100. For the all-1-invalid, all-but-1, just-2, and just-1 options, the exact RM values depended on the number of shapes transformed, as, under the complexity of two, three, and four, there were as many as four, six, and eight valid transformations on average, respectively. As one transformation was missing in the all-but-1 options, the RM values were 75 (3/4, as three out of four transformations were present), 84 (5/6), and 88 (7/8), for complexity of two, three, and four shapes, respectively. In the all-1-invalid options, it was assumed that the invalid transformations likely served as cues that these options cannot be correct. The redundant transformation was therefore scored as an RM decrement of half of the valid transformation. Thus, the resulting RM values for the all-1-invalid options under the complexity of two, three, and four were 62 (2.5/4), 75 (4.5/6), and 82 (6.5/8), respectively. For the just-2 options, on average there were two valid transformations instead of four, six, and eight expected transformations, so the resulting RM values were 50 (2/4), 33 (2/6), and 25 (2/8), respectively. Consequently, the RM values for the just-1 options equaled half of the latter values: 25 (1/4), 17 (1/6), and 13 (1/8), respectively. For the no-transformation options, RM always equaled 0. Notably, the mean RM of options was balanced across the three complexity conditions and equaled 52. In each such condition, there were always twelve visual features in total (three per shape), therefore the perceptual complexity was always constant across the test items.

Another two dependent variables were derived from the RM metric and calculated participant-wise. *Individual RM* (IRM) was the mean RM across all the 18 analogies attempted by a given participant. A person who solved all the analogies correctly scored IRM = 100, while each error decreased her or his IRM. Two people who selected the same proportion of correct options (had identical response accuracy) could still differ in their IRM if one person tended to select error options that more closely matched the correct option relationally (the all-but-1 and all-1-invalid options), while the other person processed fewer transformations, selecting error options that weakly matched the correct option, being similar perceptually to C (the just-2, just-1, and no-transformation options). The second derived variable was *error RM* (ERM) – the participant-wise mean RM calculated only for the incorrect responses. For the four people who made no error, ERM was not calculated.

The last dependent variable was response time (RT in seconds; s), analyzed both trial- and participant-wise. RT equaled the time that elapsed between the presentation of the analogy and the pressing of the accept button. Responses elapsed or shorter than 10 s were treated as errors and not included in latency analyses (there were 404 such responses).

Working memory tasks

Three variants of the complex span task, modified after Conway et al. (2005), were used to measure WMC. Each variant required the participants to memorize four, six, or eight (set size) stimuli, presented for 1.2 s apiece. Each stimulus was followed by a simple decision task to prevent the chunking of stimuli. The participants had to recall as many stimuli as they could (in the proper order) and to provide correct answers in the decision tasks. In each variant, 5 trials for each set size (in increasing order) were presented. The letter span task required the participants to memorize letters while indicating with a mouse button whether intermittent simple arithmetical equations were either correct or not. The digit span consisted of memorizing digits, and the decision task was to decide whether the 5-letter string presented after each stimulus either started or ended either with a consonant or a vowel. In the figure span task, the participants memorized geometric figures while judging colors as either light (yellow or beige) or dark (brown or navy blue). During the response procedure, as many 3 x 3 matrices as a particular set size were displayed. Each matrix contained the same set of all nine possible stimuli for a given task. The participants had to select with a mouse those stimuli that had been presented in a sequence, in the correct order. There was no time limit for responding. The dependent variable for each complex span task was the proportion of correctly selected stimuli of the 90 stimuli presented in the task. These three scores were used to compute the WMC factor using principal component analysis (PCA).

Results

There were 1980 correct responses (out of 4734; mean accuracy 41.8%). Accuracy and RT for the three complexity conditions are presented in Table 1. Response accuracy was comparable across the conditions (a drop of 1.5% from two to four shapes transformed), $F < 1$. Actually, complexity increased the proportion of transformations selected (a rise in RM by 5.2), $F(2, 4733) = 4.96$, $p = .007$, but the effect was weak, $\eta^2 = .02$. Complexity affected RT in correct trials (a rise of 4.7 s), $F(2, 1977) = 5.09$, $p = .006$, $\eta^2 = .01$, but not in incorrect trials (a non-significant difference of 2.8 s), $F(2, 2347) = 1.96$, $p = .140$. Incorrect trials were faster by 3.6 s than correct ones, $F(1, 4328) = 17.21$, $p < .001$, $\eta^2 = .01$.

Complexity:	Two shapes	Three shapes	Four shapes
Accuracy (%) 95% CI	42.6 [40.2, 45.1]	41.7 [39.3, 44.1]	41.1 [38.7, 43.5]
Relational match 95% CI	72.7 [70.9, 74.4]	73.7 [72.0, 75.4]	76.4 [74.7, 78.2]
Correct RT (s) 95% CI	59.4 [57.4, 61.5]	61.3 [59.3, 63.3]	64.1 [62.0, 66.2]
Incorrect RT (s) 95% CI	56.3 [54.2, 58.4]	58.5 [56.4, 60.6]	59.1 [57.0, 62.2]

Table 1: Accuracy, RM, and RT as a function of complexity

Fig. 3 presents the distribution of error responses (grey bars). It was collapsed over the number of shapes transformed, following the fact that the differences in RM between the conditions were negligible. The distribution differed significantly from the even distribution, $\chi^2(4) = 12.34$, $p = .014$. The proportion of each type of error option was a linear function of its average RM value, $r(3) = .969$ [.598, .998], $p = .007$. Crucially, also RT (black line) increased in a linear way as a function of the RM of error response options, $r(3) = .938$ [.323, .996], $p = .018$.

When RT was split into the three complexity conditions, and the three cases of correct responses (i.e., RM = 100 for the two, three, and four shapes transformed, respectively) were included, there was a strong positive correlation between the resulting 18 RM values and the respective 18 RT values, $r(16) = .838$ [.610, .938], $p < .001$ (see Fig. 4).

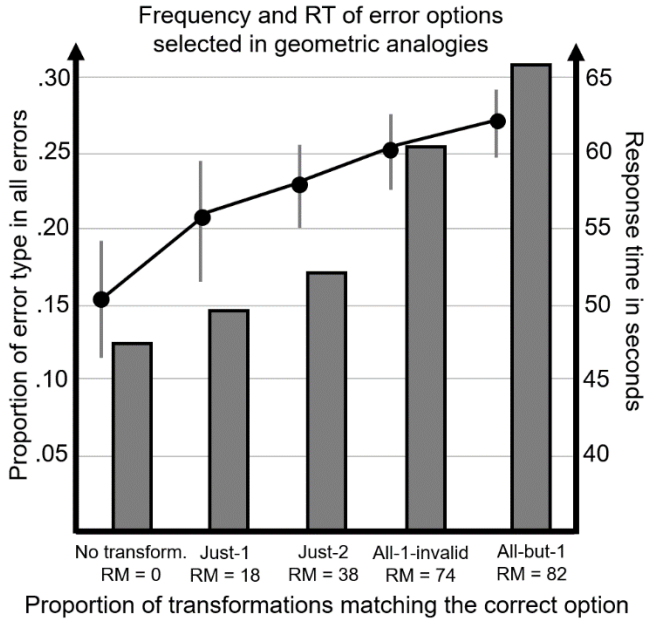

Figure 3: Proportion of each error option in all error options selected (grey bars) and a respective mean RT (black dots). Vertical lines = 95% CI. RM = the mean relational match.

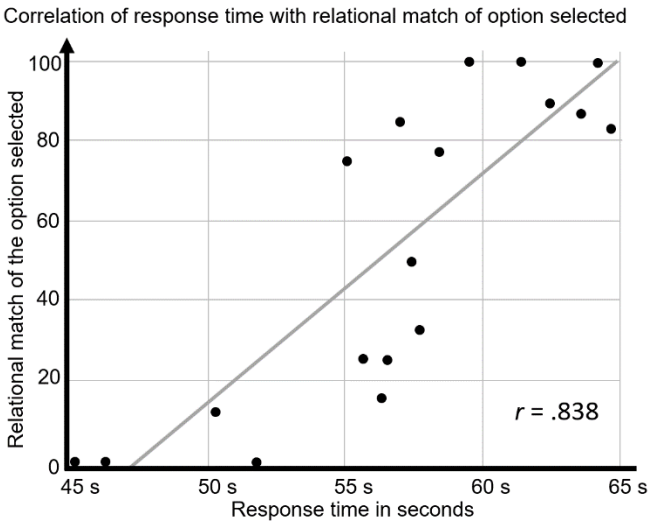

Figure 4: Correlation between the RT and the RM values of options across the three complexity conditions.

The WMC factor was defined as the principal component yielded by the PCA that was applied to the three WM task scores, Eigenvalue = 2.21, 73.7% variance explained, each factor loading = .86. The WMC factor significantly predicted response accuracy, $r(263) = .479$ [.380, .567], $p < .001$. Crucially, WMC correlated positively with individual RM (IRM), $r(263) = .535$ [.443, .616], as well as with error RM (ERM), $r(259) = .393$ [.285, .491] (see Fig. 5), both $p < .001$.

Figure 5: Correlation between the WMC factor values and the individual error RM (ERM) values for 259 participants who committed at least one error on geometric analogies.

Discussion

The results indicated that the amount of relational match of error options strongly affected their selection rate. When committing errors, people tended (in 56% cases) to select error options that included all geometric transformations needed, except one single transformation that was either

missing or wrongly substituted with another transformation. This suggests that when committing an error, the participants most frequently had already developed a relatively complex representation of analogy, but failed to complete the mapping of all its elements. Following perceptual similarity (ignoring the relations and selecting the just-1, just-2, and even the no-transformation option) was the case observed less frequently.

However, this data pattern was strongly modulated by the participants' WMC. The regression line in Fig. 5 indicates that in error trials the participants with the highest WMC (around 2 *SD*s above the mean) were able to map on average the two thirds of transformations required; in such trials the participants displaying around 2 *SD*s below the mean WMC achieved to map only less than half of transformations required. These individual differences suggest that even though relational reasoning was a main way of coping with the analogy task, certain participants might have used some simplified strategies, for instance might have been relying on perceptual similarity when selecting response options.

Less complex analogies (two shapes transformed) were solved with a comparable accuracy as was observed for more complex analogies. The proportion of transformations correctly mapped was even slightly alleviated in the latter analogies. It seems that in the present kind of sequential task (with transformations most likely mapped one by one), and with ample time allowed, relational complexity of analogies not only was not affecting their difficulty, but even a larger number of relations that bound the source and the target to some extent facilitated recognizing of relations by the participants (for a similar result see Livins & Doumas, 2015).

Methodologically, the present study suggests that the characteristics of the error response options in the multiple-choice analogy tests can significantly affect the test validity. Including in the test primarily the options that closely match relationally the correct option (i.e., miss few its elements) might increase the test's difficulty, so that even highly performing people could face problems while solving it. Such a test would validly tap into recognition and mapping of the key relations, but would be barely solvable by less performing people. By contrast, increasing the relational distance between the correct and the error options might increase mean accuracy, but could effect in testing primarily a strategy applied (i.e., either mapping or some heuristic) instead of the effectiveness of reasoning itself. People relying on reasoning would easily notice that error options cannot be correct, whereas people using heuristic strategies, such as perceptual similarity, might be prone to perceptually (but not relationally) matching options. Therefore, the specific design of error options is a crucial decision affecting the validity of any multiple-choice analogical reasoning test, and it should depend on the specific research objectives. Although the present geometric analogy test included relatively univocal features and transformations, and might not necessarily generalize onto the tests in which rules need to be discovered (e.g., Raven's Progresssive Matrices), or semantics play a role, this study nevertheless sheds some light on reasoning processes that lead to committing errors in analogy making.

References

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. *Behavioural and Brain Sciences, 24,* 87-114.

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.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10,* 277–300.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Hosenfeld, B, van der Maas, H. L. J., & van den Boom, D. (1997). Indicators of discontinuous change in the development of analogical reasoning. *Journal of Experimental Child Psychology, 64,* 367-395.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., … Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia, 46,* 2020–2032.

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.

Livins, K. A., & Doumas, L. A. A. (2015). Recognising relations: What can be learned from considering complexity? *Thinking & Reasoning, 21,* 251-264.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometry analogy problems through two-stage analogical mapping. *Cognitive Science, 33,* 1192-1231.

Novick, L.R., & Tversky, B. (1987). Cognitive constraints on ordering operations: The case of geometric analogies. *Journal of Experimental Psychology: General, 116,* 50-67.

Primi, R. (2000). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30,* 41-70.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94,* 249–273.

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development, 38,* 10-26.