PSIHOLOGIJA, 2018, Vol. 51(3), 351–375

UDC 159.95.072.59
DOI: https://doi.org/10.2298/PSI161031003Z

Multiple solutions test
Part I: Development and psychometric evaluation

Marko Živanović[1], Jovana Bjekić[2], and Goran Opačić[1]
[1]*Department of Psychology, Faculty of Philosophy, University of Belgrade, Serbia*
[2]*Institute for Medical Research, University of Belgrade, Serbia*

As people outside the context of testing seldom find themselves in situations where they are presented with limited options and a single correct answer, with all others being equally wrong, a modification of traditional intelligence tests (in terms of increasing its flexibility), can potentially provide a more comprehensive and a more valid measure of intelligence. Therefore, the aim of this study is the development and psychometric evaluation of the figural reasoning test in the form of matrices with multiple solutions. Unlike conventional intelligence tests, in this test the subjects are faced with more than one task, i.e., to detect: 1) *the best* solution – a figure that completes a given matrix best; 2) *the second-best* solution – a figure that would complete the matrix in the best way if *the best* answer was absent; 3) *the least accurate* option – a figure that completes the given matrix in the least accurate way. In the process of test development, an initial set of 80 items was designed and administrated to a sample of 41 participants, with the goal of gaining insight into the quality and the need for adjustments of the initial item pool. Psychometric characteristics of the instrument consisting of 74 items with three types of tasks have been evaluated on a sample of 263 participants, after which the short version of the instrument is proposed. All three tasks within the test and test as a whole have shown good internal psychometric properties ($\alpha_{\textit{the best}} = .92$, $\alpha_{\textit{the second-best}} = .90$, $\alpha_{\textit{the least accurate}} = .87$; $\alpha_{\textit{full-scale}} = .95$) offering a possibility of reliable measurement of intelligence with a brother scope.

Key words: Multiple solutions test (MST), figural reasoning, fluid ability (*Gf*), test development, psychometric properties

Highlights:

- Unlike conventional matrices Multiple solutions test faces a participant with three tasks within each item: to detect the best, the second-best, and the least accurate solution.
- A multi-stage process of test development is presented in detail.
- The instrument demonstrated good internal psychometric properties.
- Potential of the instrument and its alternative tasks in measuring intelligence in a more flexible and broader manner is discussed.

Corresponding author: marko.zivanovic@f.bg.ac.rs

Acknowledgement. The study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (project number 179018); The second author was supported by project number 175012.

In recent decades, the research of intelligence has greatly contributed to our understanding of the nature and structure of human cognitive abilities. Up to date, the consensual hierarchical model of the human intellect, Catell-Horn-Carroll's model (CHC), presents the theoretical basis for most contemporary ability tests (McGrew, 2009; McGrew & Wendling, 2010). This model represents an elaboration and an extension of Cattell-Horn's model of fluid (*Gf*) and crystallized (*Gc*) intelligence and Carroll's Three-stratum model (Carroll, 1993, 1997, 2005; Horn & Cattell, 1966; McGrew, 2009). According to CHC model, factors of intellectual abilities are organized into three levels of generality (Carroll, 1993, 1997, 2005; McGrew, 2009; McGrew, & Wendling, 2010). On the first level, numerous narrow or specific abilities are located. On the second level, there are around ten group-level factors, most of which were originally defined by Cattell and Horn, among which *Gf* and *Gc*. *Gc* reflects individual differences in knowledge and the depth of language knowledge, concepts of culture. It's acquired in the process of education, by the accumulation of experience, and primarily reflects the verbal knowledge and skills as well as declarative knowledge in various fields. On the other hand, *Gf* represents the capacity to solve new, complex problems using inductive and deductive operations and reasoning. This ability includes processes of comprehending relations, usually inferring from incomplete information and thus represents one of the most central components of intelligence (Carroll, 1993, 1997, 2005). These processes are at work whenever the perception of complex relations is present (Carroll, 1993, 1997, 2005; Cattell, 1987). Abilities that underlie *Gf* include complex mental operations such as: perceiving relations, extrapolation, concept formation, generating and testing hypotheses, etc. The general factor of intelligence (*G*) is positioned on the third level. *G* underlines and subsumes all of the complex higher-order cognitive processes (Carroll, 1993, 1997, 2005).

Measures of fluid abilities

Although, according to the CHC theory *Gf* is located in the middle stratum of the hierarchical structure of the intellect (Carroll, 1993, 1997, 2005; McGrew, 2009; McGrew & Wendling, 2010), and represents one of the broad factors subordinated to *G*, some authors (e.g., Gustafsson, 1984) believe that *Gf* as described by Cattell and Horn (Horn & Cattell, 1966) actually represents Spearman's *G* (Spearman, 1904, 1927), and therefore consider it more fundamental than the other second-order-factors of intelligence[1] (Gustafsson, 1984, 1988, 1994, 2002; Jensen, 1982; Jensen, 1998).

Some of the tests that are considered to be good measures of *Gf* are tests that engage inductive reasoning, concept formation, visual conceptualization, efficiency in using of strategies in solving problems, etc. (Horn, 1988). As

1 For alternative view and evidence against this claim see e.g., Süß & Beauducel (2015).

figural reasoning constitutes one of the defining abilities that underline *Gf* (Horn, 1979), abstract geometric stimuli have found wide application in the construction of items aimed at measuring *Gf*. The material commonly used to measure *Gf* is either equally known or equally unknown to everyone (Cattell, 1987). Since figural reasoning tasks tend to be highly saturated by *G* due to a high level of complexity and abstraction (Snow, Kyllonen, & Marshalek, 1984) some authors consider these types of tests some of the best measures of *Gf*, as well as G^2 (Jensen, 1998).

One of the most successful and most commonly used operationalizations of this ability are matrices. In these tests, the subjects are faced with a matrix consisting of figures that are arranged in rows and columns by a set of rules. One element in a matrix is missing, and it is to be found among offered options. Participant's task is to pick among the several presented elements and select a figure which completes the matrix, guided by the rules that apply to a given matrix. Tasks of this type are usually not time-limited (or have relatively liberal time constraints), and represent pure power tests (Jensen, 1998).

One of the most widely used intelligence tests consisting exclusively of matrices is Raven's Progressive Matrices test (RPM) (Raven, 1938; Raven & Court, 1998). This test is designed as an operationalization of Spearman's education of relations and correlates (Penrose & Raven, 1936; Raven, 1938; Spearman, 1904, 1927). Some of the reasons for the wide application of this test is the simplicity of setting and evaluation, the possibility of individual or group administration, applicability to a wide age-range, and since it is a nonverbal test it has the possibility of administration to the different language communities without special adaptations. Consequently, RPM is one of the most widely used measures of intelligence both in research and in practice (Mackintosh, 1998; Raven, 2000). Many authors consider it to be one of the best non-verbal test and a pure measure of *G* (Carroll, 1993; Jensen, 1998; Spearman, 1946; Vernon & Parry, 1949). The validity of highly *G*-saturated tests, such as matrices in the prediction of a variety of relevant criteria has been demonstrated in a number of studies (Gottfredson, 1997; Jensen, 1998; Kuncel, Hezlett, & Ones, 2004; Salgado et al., 2003; Schmidt & Hunter, 1998). Additionally, as a subscale, matrices are included in a number of widely used batteries for intelligence assessment.

Criticism of traditional intelligence tests

Although conventional intelligence tests show a wide practical validity through prediction of a number of relevant external criteria, even higher than any other psychological construct (Gottfredson, 1997; Jensen, 1998; Kuncel et al., 2004; Salgado et al., 2003; Schmidt & Hunter, 1998), some authors still question their validity. Namely, a number of authors who emphasize contextual importance raised the question of whether the traditional items in intelligence

2 For the evidence against this stand see e.g., Gignac (2015).

tests reflect one's intellectual capacity that can be related to success in solving real-life problems (Ceci, 1990; Gardner, 1993; Sternberg, 1985; Sternberg & Wagner, 1986).

One of the frequent criticisms regarding traditional intelligence tests is that they present a person with a problem offering a *single* correct answer, as well as having just one way of getting to that answer (Sternberg & Wagner, 1993). In such a situation, a person is unable to use some of his/her potential preferences or styles that have shown to be successful in his/hers real-life problem-solving, or to deal with a problem in an alternative manner, which stems from the knowledge of their own strengths and weaknesses (Sternberg, 1999). So one can imagine a situation where a person who underperforms on a traditional intelligence test, yet have no difficulties in solving problem situations in real-life, due to a possibility of dealing with a problem in an alternative way, which, in the context of the standard intelligence tests, is not possible. Such alternative solutions to a problem should certainly be considered equally successful adaptations.

Thus, as the people outside of the context of tests, seldom find themselves in situations where limited options are presented, one being accurate, and all other equally "wrong", a modification of traditional tests, in terms of increasing their flexibility can potentially provide a more comprehensive and a more valid measure of intelligence.

Modification of the matrices

One of the tests that maintained the basic formal properties of traditional intelligence tests, but introduced a modification of the items in order to bring the test closer to real-life problem solving is the *Test višestrukih rešenja* (TVR) (Bujas, Bartolović, & Vodanović, 1967). Authors emphasized the role of flexibility in intelligent behavior and tried to make the test content, more complex and closer to real-life problem-solving. Here the flexibility is understood as the ability to 'identify an imbalance in different situations' and thus they have defined it as 'sensitivity' to a problem (Bujas, 1966). TVR represents a modification of the standard matrices with one type of solution. Namely, subjects are not supposed to find a single correct solution, but rather *three* requirements (tasks) are set before them: finding the *correct* solution (this task is identical to standard matrices), finding the solution that is *approximately correct*, which does not meet all the criteria for the correct one), and to find the figure that deviates from the correct one the most (or *the worst option*) (Bujas, 1966; Bujas et al., 1967). In that way, the sensitivity of traditional matrices is largely increased. According to authors, intelligence tests generally assume that there is only one correct solution and that all the other options except that one are equally "wrong". In real-life problem-solving, a person rarely faces only one solution that is the best for everyone. The best solution depends, both on the specific situation, i.e., problem in question, and what one wants to achieve, as well as the means available to achieve it, therefore, people of equal intelligence could behave differently in actual problem situations (Bujas et al., 1967).

Despite the fact that the basic premise of Bujas et al. (1967) represents a framework which integrates some of the important ideas of the contextual paradigm and rigorous psychometric criteria that are set before intelligence tests, the operationalization of this idea through TVR faces several important problems. First of all, when analyzing solutions in this test, it is difficult to clearly explicate the rules by which one solves *approximately correct* and *the worst* task. *The approximately correct* option is usually the option that's most alike (most similar) to *the best* one. This raises the question of whether perceptual discrimination is, in fact, *the only* process engaged in solving *the approximately correct* task. If this is the case, it can be argued that this task faces a construct validity problem, i.e., which aspects of intelligence are measured by it and the extent to which these aspects differ from the abilities needed to solve the task in a standard form. Likewise, solving *the worst option* task faces the same problem, and it seems to rely on neither the general ability nor fluid abilities. In addition to being the one that by default is not sharing a single relevant characteristic with the *correct* and *the approximately correct* choice, *the worst option* also represents *perceptually* the least adequate option. Therefore, it's questionable that solving these two tasks primarily relies on the abilities that are central to intelligence.

The present study

One of the potential shortcomings of intelligence tests may be the formal distance of task demands from those required in real-life problem-solving. In fact, in real-life, problem situations rarely require only one "correct" solution (Bujas et al., 1967; Sternberg & Wagner, 1993). On the contrary, people are more often faced with a problem that may involve finding "the second-best" option when the perfect one is unattainable or at least detecting the least desirable alternative. Therefore, the differentiation between options which are more or less accurate presumably more closely resembles real-life problem solving, and abilities involved in solving them. Hence, while increasing the flexibility of the intelligence test, with emphasis on the measurement of cognitive flexibility, and preserving the objectivity of the test method, it can be assumed that the demands imposed by the test are brought closer to the demands of real-life problem-solving.

The novelties reflected in the alternative tasks incorporated in standard matrices introduced by Bujas and colleagues (Bujas, 1966; Bujas et al., 1967) and which will be addressed and elaborated within this paper can be viewed as an extension of the proven measure of *Gf*. The alternative tasks could be interpreted in terms of additional demands that are imposed on building blocks of higher-order cognition, i.e., executive functions. Namely, it can be assumed that processes captured by these alternative tasks are more closely linked to executive control, i.e., initiation of goal-directed behavior, inhibition of competing stimuli characteristics, selection of relevant and suppression of irrelevant aspects of a task, flexible shifting of problem-solving strategies, planning, monitoring, and evaluation of behavior directed at problem-solving. Some of the executive

processes are proven to be assessed by conventual intelligence tests, i.e., working memory and updating (Ackerman, Beier, & Boyle, 2005; Chuderski, Taraday, Nęcka, & Smoleń, 2012; Colom, 2004; Colom, Abad, Rebollo, & Shih, 2005; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Friedman et al., 2006; Kane, Hambrick, & Conway, 2005; Kane et al., 2004; Kyllonen & Christal, 1990; Martínez et al., 2011; Oberauer, Süβ, Wilhelm, & Wittmann, 2008), while some of the important supervisory processes are not sufficiently covered by the standard intelligence tests (see Friedman et al., 2006) in spite of their high relevance in everyday functioning and validity in the prediction of a variety of relevant real-life criteria (see Diamond, 2013). Thus, enriching one of the *G*-most-central tests (Carroll, 1993; Jensen, 1998; Snow et al., 1984; Spearman, 1946; Vernon & Parry, 1949) by introducing measures which would potentially broaden its object of measurement seems to have a potential to provide more comprehensive and more valid assessment of one's core intellectual capacities. So complementing proven measure of the fluid abilities (finding *the best* solution) by alternative tasks (finding *the second-best* and *the least accurate* solution) would potentially provide a relevant proxy for one's ability to adequately employ his/hers operative capacities in a flexible manner.

The aim of this study is development and psychometric evaluation of the figural reasoning test in the form of matrices with multiple solutions based on Bujas et al. (1967). The development of this test represents an attempt to renew and enhance an important idea that combines the psychometric and the contextual paradigm, and their main advantages: an effective way of assessing intelligence and bringing intelligence tests closer to real-life problem-solving, therefore making them more valid.

In this paper, the process of instrument development will be presented in detail, followed by a psychometric evaluation of the test, while succeeding article (Part II in this issue of Psihologija) reports on construct and predictive validity of the instrument (Živanović, Bjekić, & Opačić, 2018).

Test development

This section presents steps in the development of the instrument along with a brief description of its formal and content characteristics and alternative solutions. This section also presents results of the preliminary evaluation of the instrument.

Test description

The content and format of items and tasks. The format of the items is based on traditional matrices for intelligence assessment. Items consist of a matrix 2x2, a matrix 3x3, and an array of figures in a one-row matrix. In all items, entries of the matrices/arrays contain figures that alter through rows and columns, following a set of rules that apply to a given matrix. Figures used in

the test are designed to consist of basic geometric shapes. In each item one entry of the matrix is blank, and the participant needs to complete the matrix choosing among the presented options. The entry of interest can be found anywhere in the matrix. All the items are closed-ended, with six options offered. Unlike conventional matrices, in addition to identifying the correct figure, a person needs to respond to two more requirements. Thus, in every item participant needs to solve three types of tasks, i.e., to solve for: *the best solution* – among multiple options presented, one needs to find the one that completes the matrix in the best way, following all the rules that apply to the matrix (identical to standard matrices); *the second-best option* – among available options, one needs to realize which figure would complete the matrix in the best way if the correct answer were not offered, i.e., the figure which deviates from the matrix rules the least, in comparison to all other choices except *the best* one; *the least accurate option* – among available options, the participant needs to detect the figure that completes the matrix in the least accurate way, i.e., the figure which, in comparison to other choices, deviates from the rules that apply to the matrix the most.

Although these three types of tasks have been based on the test of Bujas et al. (Bujas et al., 1967), and the two tests share almost all formal characteristics, the items themselves differ in several important aspects. Firstly, in TVR finding *the correct* solution nearly automatically leads to finding *the approximately correct* one, given that between these two tasks there is a high level of visual similarity. Therefore, finding *the approximately correct* option in TVR could easily be called "detecting the most-similar-to-the-best figure". Consequently, an obvious perceptual similarity between the figures that meet the first two requirements, leads to the fact that respondents can find *the approximately correct* option simply by comparing the remaining available options with *the correct* one. In contrast, the items of MST are designed in such a way that *the best* and *the second-best* option share the rules that apply to a given matrix, but *the second-best* option partially deviates from the given rules, whereas the level of perceptual deviation/similarity from/to *the best* one, is not crucial.

Furthermore, in TVR, *the worst option* is the one the features of which differs the most, both from the figures in a given matrix, as well as from all other choices offered. In MST *the least accurate* option is designed to share some of the crucial features with the figures from the matrix and available choices, but this option is designed to deviate from all or most of the rules that apply to a given matrix. Finally, in TVR, after the detection of *the correct* option, the focus shifts to a successive comparison of the remaining options and detecting remaining solutions. Conversely, in MST one has to figure out all three solutions exclusively relying on the rules that apply to the matrix. Thus, here the primary focus is on the rules that have to be detected, understood, and followed as the key concept of solving a matrix (Carpenter, Just, & Shell, 1990), while mutual comparison between other available choices in order to detect the adequate ones represents the second step in finding the remaining solutions.

Designing adequate solutions and distractors

Following the setting of a basic structure of the item and rules that apply to a matrix, the initial step in developing every item was to design *the best* option. *The best* option, just like in traditional intelligence tests, is the figure that completes the matrix in the best way, following all the rules that apply to the alternation of elements through rows and columns of a matrix. An example of the item from the test is shown in Figure 1.

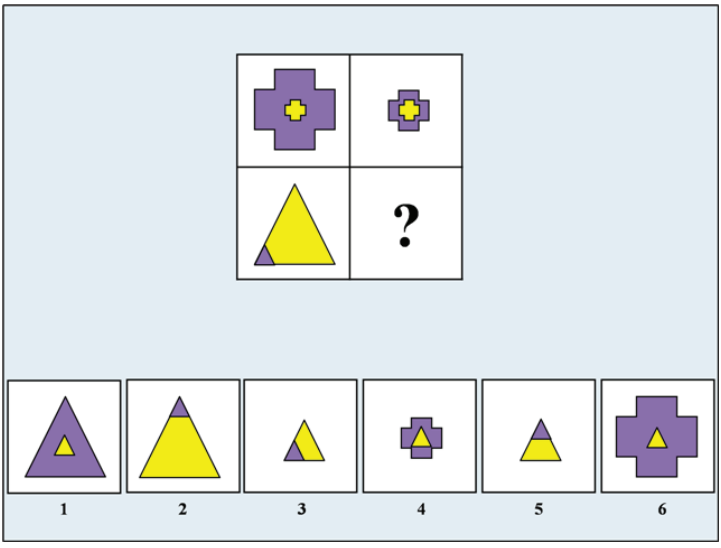

Figure 1. Example of test's item.

As can be seen, the problem and the task (to find a figure that adequately complements the matrix) are identical to the standard test matrices. *The best* solution to the problem shown in the figure is option 3. This option follows the rules that apply to the alternation of elements through rows and columns of the matrix and therefore represents the "ideal" solution. In each task, there is only one figure that adequately completes the matrix. Other available options, at this stage of the problem-solving, serve as distractors. At the same time, each of these options is, more or less correct having in mind the rules that apply to the matrix. In other words, for each option it is possible to quantify its degree of deviation from the rules following the number of deviations, thus ranking them according to the relevant criteria met. The possibility of grading the accuracy of the options in this way has served as the basis for the development of adequate solutions to the two other requirements – *the second-best* and *the least accurate* option.

The second-best option. *The second-best* solution is operationally defined as a figure which deviates from the matrix rules the least in comparison to all other choices, except from *the best* one, i.e., a figure that would complete the

matrix the best if the "ideal" solution were absent among the options available. So if in a given task there are, for example, three relevant rules (rules that lead to the "ideal" solution) *the second-best* solution is the one that meets two out of three. In the task depicted in Figure 1, *the second-best* option is the figure 5. This option follows the alternation of figures through rows and columns of the matrix, i.e., it has the appropriate shape, adequate color, position of the inner figure, etc., however, the position of the inner figure is distorted. Observing remaining options, it is clear that figure 5 is the option that is closest to the one that would follow all the relevant rules that apply to the matrix. In other words, other available options deviate from the rules in more than one aspect. In the case of the example presented in Figure 1, it is evident that *the second-best* and *the best* solution are somewhat perceptually similar. In easier items, in most cases, this is inevitable due to the insufficient number of rules that can be varied, and it is often impossible to avoid similarity of perceptually dominant dimensions.

However, in more difficult items, or in items where more complex rules are applied perceptual similarity between these options is not an issue. An example of such an item is shown in Figure 2.

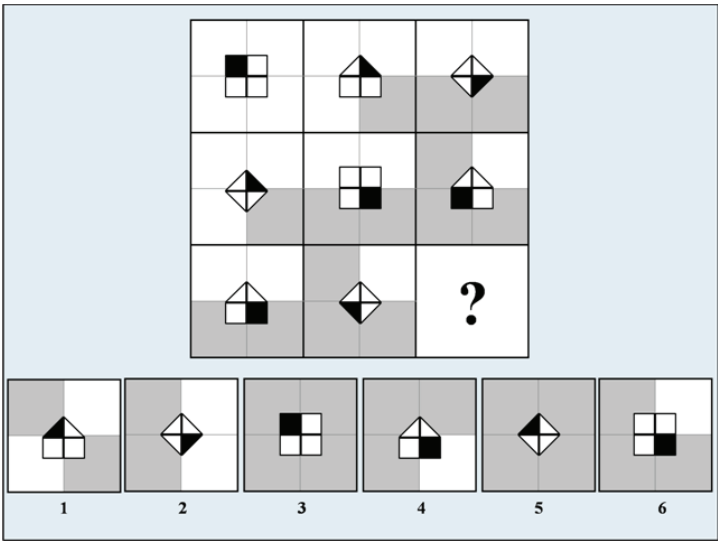

Figure 2. Example of test's item.

In this case, *the best* solution (number 3) displays much lower perceptual similarity to *the second-best* one (number 5), then to one of the distractors (number 6). However, figuring out the rules of figure alternation, it is clear that figure 5 is the one that follows most of the rules that apply to this matrix. Thus, in the absence of the "ideal" solution, its deviation from the matrix rules is the least in comparison to other available options. In order to find an adequate solution to this task, one has to extract a combination of relevant aspects which

make this option *the second-best*, whereby s/he has to ignore aspects by which this figure deviates from the rules.

The least accurate option. *The least accurate* option is operationally defined as the one that, in comparison to all other available options, deviates from relevant rules the most. This option can be a figure that doesn't meet *any* relevant criteria as in Figures 1 (6) and 2 (2) or a figure that meets *some* of the relevant rules. In all cases, the number of rules that *the least accurate* option follows is lesser than the number of relevant aspects that are met by other available options. As in the case of *the best* and *the second-best* option, for each task, there is only one adequate solution. As noted before, in most of the cases *the least accurate* option is not the one that is perceptually distorted the most, nor the one that perceptually stands out from the rest of the figures. In order to adequately solve this task, one has to combine the relevant aspects of the task simultaneously and detect the option which compared to all other, deviates from the rules the most.

Distractors and guessing. The designing of the distractors was carried out simultaneously with the designing of adequate responses for each task. During the distractor designing, we specifically paid attention to the notion that every adequate solution has to have about an equal number of competitive alternatives. As *the best* and *the second-best* option, are competitors in the tasks for solving for *the best*, and *the second-best* option, in most of the items, one more figure was designed to further distract detecting the adequate solution to these two tasks. The other two distractors are made with the aim to compete for an adequate solution with the tasks of finding *the least accurate* option. So in most of the items, meeting each of the tasks involved a choice between three competing options. For example, in the case of the item shown in Figure 1, Figures 3, 5, and 6, are competitors for *the best* and *the second-best* option, while Figures 1, 2, and 4 are competitors for *the least accurate* option.

This decision seemed to be the most appropriate given the relatively high probability of guessing due to the ratio of the available options and the number of tasks. Specifically, since the items are designed so that the participant would simultaneously provide answers to three tasks, the degree of differentiation between the options offered had to be diverse (with a clearly quantified number of distinctive features – which is uncommon in tests). Of course, these precautionary measures are instrumental under the assumption that respondents actually recognize the similarity between competing options. Since it can be assumed that the order of solving the item is: finding *the best* solution first, followed by identifying *the second-best*, and lastly, finding *the least accurate*, the decision to place *the best* and *the second-best* option in a set of three mutually competitive options seemed to be the most appropriate.

In other words, the probability of guessing each of the solutions within each item can be considered to depend on: a) the order responding to the tasks and b) the accuracy of solving previous tasks. Since it can be assumed that the

order of solving tasks is constant and ordered from detecting *the best* solution followed by the other two, probability of guessing increases from *the best* to *the least accurate* one. However, since the probability of guessing increases along with the accuracy of solving previous tasks within the same item, it can be assumed that the more successful a person is in solving previous tasks, his/her probability of "guessing" increases[3]. Thus, it can be assumed that the potential guessing does not represent a noise in whole, yet, to a large extent reflects the ability measured.

The positioning of adequate solutions for *the best*, *the second-best*, and *the least accurate* figures are counterbalanced so that each adequate solution is positioned on each of the six places among options available, approximately equal number of times. Positions of adequate options for *the best*, *the second-best*, and *the least accurate* were then pre-randomized across the items.

Item difficulty

During the test development, item difficulty of individual tasks was established through a complexity of rules being varied, i.e., the number of relevant aspects that are to be comprehended and followed in order to reach the adequate solution (see Carpenter et al., 1990), as well as through the similarity of competing options. So in easier items, the number of aspects that are to be figured out in order to solve them is relatively small. As the number of relevant aspects of the item to be comprehended increases, the difficulty of the task increases as well. Similarly, in the easier items resemblance of the adequate option with other available choices is lesser than in the more difficult tasks, making discrimination easier. During test development, steps were undertaken to ensure that the format of the item (matrix type and size) is not correlated with item's difficulty, so that the same number of arrays of figures, a 2x2 matrix, and a 3x3 matrices are approximately equally complex. This achieves that the task difficulty depends solely on the content features of the item, and not its format.

Since solving *the second-best* and *the least accurate* task presumably require higher levels of inference in comparison to *the best* one, it can be expected that these two tasks are somewhat more difficult. Namely, it can be assumed that these two tasks put more load on one's executive control – one has to keep in mind all the matrix rules and simultaneously shift attention from one to another,

3 The probability of guessing the adequate solution in solving for *the best* option is .17. The probability of adequate option guessing in solving for *the second-best* option depends on the success in solving previous one. Namely, if in the previous task adequate figure is selected probability of guessing *the second-best* option is .20. If in the previous task *the best* solution has not been found, and any other option except *the second-best* was chosen, the probability remains .20. But if the person has chosen *the second-best* option as being *the best* one, the probability of guessing and finding an adequate option for *the second-best* task is zero. Finally, in *the least accurate* task, participants who have properly solved for previous two solutions, have a probability of guessing of .25, as well as participants who have made a mistake, but in none of the previous tasks have not chosen adequate solution for *the least accurate* task, as being *the best* or *the second-best*.

inhibiting irrelevant yet salient characteristics of the figures offered, while at the same time searching for the most adequate solution among the options available. However, since respondents provide the answers to all three requests within each item, it was taken into account that individual task's difficulty within each item is relatively equal, i.e., that the difference in difficulty between the same tasks in different items is larger than the difference in task's difficulty within items.

In order to rank the items by means of their difficulty in the preliminary test, three independent raters, rated the difficulty of each of the three tasks, within each item. Raters had an insight into the correct solutions and the rules that are to be followed. In this way, it is avoided that raters give judgments on the item difficulty on the basis of their own success in solving them. Raters are instructed to evaluate the complexity of detecting the rules by which the tasks are solved and finding the appropriate solutions following given rules. Ratings were given for each task within each item separately, which resulted in three ratings per item. The ratings were then averaged across tasks, and based on them the items were arranged in ascending order of difficulty.

Timing

The test was designed with the intention of administering it within a non-restrictive time limit. In other words, the MST was administrated so that the subjects' performance depends on his/her abilities regardless of their speed of processing and the amount of time available. There are several reasons for arguing that non-speeded tests are better, or at least, more useful measures of general ability. Performance on speeded tests is significantly influenced by various noncognitive factors, such as test anxiety or different personality traits (Ackerman & Heggestad, 1997). Furthermore, a restrictive time limit leads to a lesser number of items solved and a large number of "skipped" ones, which have an adverse effect on test validity (Lu & Sireci, 2007). Finally, non-speeded tests show better predictive validity than speeded ones (Denis & Gilbert, 2012).

Preliminary study

The aim of the preliminary study was to provide an insight into the quality of the test's items, i.e., properties of individual tasks, and the test as a whole. More specifically, at this stage of instrument development, we wanted to establish informed guidelines for the modification of certain items/tasks, so as to exclude poor and confusing items/tasks and to assess the need for the development of new items. Overall, we aimed to empirically examine every item, both quantitatively and qualitatively, before administering the final version of the test. Despite the fact that during the design of items special attention was paid to the notion that there are no alternative ways or strategies for solving the items or items that have more than one adequate solution (per task), at this stage, items are tested on a sample of participants who have not previously had an experience in solving items in questions.

Thus, at this stage, the focus was primarily on the poor items, i.e. items that one could find ambiguous, and confusing, and their modifications. Furthermore, preliminary testing was used to establish an empirically based order of items in the test by means of their difficulty before administrating the test to a larger sample. Finally, for practical reasons, the information on the time needed for test completion, was obtained in this phase of research.

Method

Participants. The sample for the pilot study consisted of 41 volunteers, aged 22 to 32 years (M = 26.54, SD = 2.72), 15 men and 36 women, most of them university-educated individuals/students.

Instrument. At this stage, the instrument consisted of a total of 80 items that were designed in the development process. All the items had six available options offered, out of which the only one within each task was the correct one.

Procedure. The instrument was individually administrated to participants. After the general instructions, in the practice trial participants were familiarized with the items that were to be solved, as well as three types of requirements that will be set before them. Upon completion of the practice section, participants were asked to start the test. For each item, participants provided their answers indicating the number of the figure which they think is *the best*, *the second-best* and *the least accurate* solution. The test is administrated without time limit, yet this information was collected for practical reasons.

After the data was collected, an interview was conducted with 5 participants-volunteers, who were willing to re-solve the test, but to try to verbalize their thoughts and to explicate the strategy of solving each item, consider alternative strategies, indicate options that they find ambiguous, etc. In this way, we aimed to detect which aspects of the items and tasks pose a problem to people in finding adequate solutions. In other words, the interview served as a method for detection of ambiguous items/options, so that they could be modified/refined.

Results

As presented in Table 1, three types of tasks, in general, have shown relatively diverse difficulties – participants were detecting *the best* solution with less difficulty than *the second-best* ($t(40) = 17.122$, $p < .01$), and *the least accurate* ($t(40) = 17.450$, $p < .01$), but were equally successful in detecting *the second-best* as they were in detecting *the least accurate* option. None of the respondents achieved a maximum score, in any of the tasks, while at the same time, all the participants were able to solve at least several items. The *K-S* indicated a normal distribution of scores in all three tasks.

Table 1
Descriptive statistic for the best, the second-best, and the least accurate solution

	M	*SD*	*Min*	*Max*	*Sk*	*Ku*	*zSk*	*zKu*	*K-S*
the best	61.59	8.73	38	78	-0.69	0.69	-1.87	0.95	0.69
the second-best	43.15	11.84	16	66	-0.07	-0.40	-0.19	-0.55	0.37
the least accurate	44.32	10.14	29	65	0.17	-0.66	0.47	-0.91	0.53

Note. *M* – mean; *SD* – standard deviation; *Min* – minimum; *Max* – maximum; *Sk* – skewness; *Ku* – kurtosis; *zSk* – standardized skewness; *zKu* – standardized kurtosis; *K-S* – Kolmogorov-Smirnov test of normality of distribution of scores; ** $p < .01$; * $p < .05$.

Cronbach's alphas for *the best*, *the second-best*, and *the least accurate* task were .883, .894, .851, respectively, while average inter-item correlations obtained were .10, .10, and .07, respectively, indicating a slightly lower homogeneity of all three tasks. However, given that the measures resulted from a very small sample, the primary focus was set on the items with extremely poor psychometric characteristics, i.e., items with very low reliability and internal validity indicators, as well as on items for which there was a marked difference in difficulty between three tasks.

The items in which at least one of the tasks had a negative or zero correlation with the total score, and had poor reliability indices were considered to be modified or excluded from the test. A total of 33 items, most of which showed relatively poor psychometric properties in a single task, proved to be potentially problematic at this stage of test evaluation.

Inspection of respondents' answers (the frequency of selecting available options) and interviews with participants showed that 19 of the previously selected items required changes. On the other hand, six items, for which modification would imply a dramatic change, were excluded from the test. Modifications made to the 19 selected items involved the change of one or more options. The modifications undertaken were focused on the target-figure (adequate response) (14 items) and/or the distractors (8 items). In total, for 6 items *the least accurate* option was modified, *the second-best* option was modified for 8 items, while *the best* option hasn't been modified in any item. For all items, modifications were content-related (changes in one or more aspects of the figure), while the formal characteristics of the modified items (dimension and type of the matrix, stimulus type, etc.), as well as the rules applied, remained unchanged.

The outcome of the preliminary study. At this stage of development, the instrument was tested on a small sample of participants, with the aim of gaining insight into the quality of items designed, and the need and guidelines for their modification. Quantitative indicators have marked 33 items, which were, on the basis of qualitative analysis either modified (19), excluded from the test (6) or remained unchanged due to the inability to detect inconsistencies in the items.

This phase resulted in a version of the instrument which consisted of 74 items. On the whole, the test showed satisfactory psychometric properties. However, it should be noted that a small number of items that were not excluded or modified at this stage showed unsatisfactory psychometric properties. Since this phase was conducted on a relatively limited sample, the stability of obtained quantitative measures can be questioned. Therefore, at this stage, far less attention was devoted to these indicators, but the focus has primarily been set on the items' content-related adequacy. More specifically, this stage ensured that no item would contain options that participants find ambiguous or confusing or items to which an alternative strategy could be applied in order to reach adequate solutions.

Psychometric evaluation

The final step of the instrument development included psychometric evaluation of the test on a much larger sample. Within this phase, psychometric properties of individual tasks, and items, as well as the instrument in general, are tested. Based on the results obtained, a shorter version of the test, comprised of the items with the best psychometric characteristics within each of the three tasks, is proposed.

Method

Participants. The sample consisted of 263 respondents, from 19 to 51 years of age (M = 22.55; SD = 3.30), 46 men and 217 women. The majority of the sample was comprised of university educated individuals who volunteered to participate in the research.

Instrument. The test consisted of 74 tasks, graded by difficulty based on the parameters obtained in the previous phase. All the items had six options offered, out of which the only one within each task was the adequate one. Participants' task was to detect three types of solutions *the best*, *the second-best*, and *the least accurate* one.

Procedure. As in the preliminary study, the instrument was administrated to participants individually and consisted of general instructions, a practice section, and the main part. Participants provided their answers marking a number which represents the figure they consider *the best*, *the second-best*, and *the least accurate* solution. The test is administrated without a time limit.

Results

The full version of the test has generally shown good psychometric properties[4]. However, in order to maximize the quality of items within each task, and in order to construct a shorter (more economical test), based on quantitative parameters obtained, a short version containing 40 items, was formed. Here it is important to emphasize the specificity of the test developed. Unlike other tests in which the subject provides a single answer to each item, in this test each item requires three answers to be made. Thus, different aspects of a single item can show different (better or worse) psychometric properties. Since different tasks share the same content, poor psychometric characteristics of a single task within the item inevitably lead to the exclusion of a whole item from the test. Therefore, when selecting items for the final version of the test we were led by *the least accurate* solution's psychometric properties as a lower limit for keeping the item

4 Cronbach's alphas for *the best*, *the second-best*, *the least accurate* solution, and the full-scale were .933, .902, .879, 952, respectively. Kaiser-Meyer-Olkin (KMO) measure of item sampling adequacy for the full-scale (.971) and first two tasks were at the satisfactory level .953, .904, whereas for *the least accurate* solution was somewhat lower (.843). Homogeneity measures (*H5*) obtained for three types of tasks, and full-scale were: .451, .420, .404, and .518. All the measures were calculated using Rtt10g macro (Knežević & Momirović, 1996).

since this task has shown somewhat lower metric properties while tightening the criteria for *the second-best* and *the best* solution[5].

After the exclusion of items that didn't meet the criteria, the data set for 40 selected items was reanalyzed. Descriptive statistics for *the best*, *the second-best* and *the least accurate* solution are displayed in Table 2. The results have shown that three types of tasks differ in terms of their difficulty, $F(1.87, 490.10) = 499.71$, $p < .01$. Bonferroni post hoc tests revealed that participants more easily solved the test for *the best* option ($p < .01$), had more trouble in finding *the second-best* one ($p < .01$), and had the most difficulties in detecting *the least accurate* option ($p < .01$, $p < .05$)[6]. In each task, few participants achieved close to maximum or a maximum score, while no participant solved less than 8% of the items within each task. The symmetry coefficients indicated a distortion of *the best* solution toward higher scores, while distributions of *the second-best* and *the least accurate* solution remained symmetrical. However, kurtosis indices pointed to the platykurtic distributions of *the second-best* and *the least accurate* solution relative to normal.

Table 2
Descriptive statistics for the best, the second-best, the least accurate solution, and the MST full-scale score

	M	*SD*	*Min*	*Max*	*Sk*	*Ku*	*zSk*	*zKu*	*K-S*
the best	27.72	8.28	5	40	-0.66	-0.50	-4.40**	-1.68	2.04**
the second-best	21.28	8.64	3	38	-0.20	-1.05	-1.32	-3.50**	1.55*
the least accurate	20.55	7.63	4	39	0.06	-0.86	0.41	-2.88**	1.08
MST full-scale	69.55	23.57	17	114	-.284	-0.97	-1.89	-3.26**	1.41*

Note. M – mean; *SD* – standard deviation; *Min* – minimum; *Max* – maximum; *Sk* – skewness; *Ku* – kurtosis; *zSk* – standardized skewness; *zKu* – standardized kurtosis; *K-S* – Kolmogorov-Smirnov test of normality of distribution of scores; ** $p < .01$; * $p < .05$.

The sampling adequacy in all three tasks, as well as the full-scale[7] achieved a satisfactory level ($KMO_{the\ best} = .958$, $KMO_{the\ second\text{-}best} = .941$, $KMO_{the\ least\ accurate} = .898$, $KMO_{full\text{-}scale} = .982$). The reliability for *the best*, *the second-best*, *the least accurate solution*, and the full-scale remained at a satisfactory level, despite the considerable reduction of the items ($\alpha_{the\ best} = .919$, $\alpha_{the\ second\text{-}best} = .905$, $\alpha_{the\ least\ accurate} = .866$, $\alpha_{full\text{-}scale} = .946$). Homogeneity measures (*H5*) for *the best*, *the second-best*, *the least accurate* solution, and the full-scale increased to .574, .549, .536, .722, respectively. Table 3 displays the ranges of the items'

5 Criteria for keeping the item in *the least accurate* task were item sampling adequacy of > .70, reliability > .25, and internal validity > .25. Criteria for keeping the item in the final test for *the second-best* solution were: item sampling adequacy > .80, reliability > .30, and internal validity > .25. Items which satisfied these criteria, showed item sampling adequacy > .85, reliability > .35, and internal validity > .25, regarding *the best* solution.

6 Gender differences in solving three types of tasks were not found ($F(1, 261) = 0.607$, $p = .437$).

7 Score on each item is calculated as a sum of three tasks. Therefore, the range of obtainable scores on each item is from 0 to 3.

sampling adequacy, reliability, and internal validity within three types of tasks, calculated by RTT10g macro (Knežević & Momirović, 1996).

Table 3
Ranges of item sampling adequacy, reliability, and internal validity within the best, the second-best, the least accurate solution, and the MST full-scale

	item sampling adequacy	reliability	internal validity	
			H	*B*
the best	.83–.98	.21 – .53	.25 – .68	.29 – .67
the second-best	.76–.97	.16 – .52	.24 – .66	.27 – .64
the least accurate	.73–.96	.15 – .37	.23 – .60	.26 – .58
MST full-scale	.95–.99	.25 – .56	.41 – .73	.42 – .72

Within all three tasks, there were a small number of items that had a slightly lower, but still fair sampling adequacy. The majority of items within all three tasks showed satisfactory reliability. Likewise, all the items in all three tasks expressed good internal validity. None of the items within each task have shown zero or negative correlation with the object of measurement.

In order to further examine the items' properties within three tasks and features of tasks as a whole, the Principal component analysis was performed (Table 4).

Table 4
Results of the Principal component analysis for the best, the second-best, the least accurate solution, and MST full-scale scores

comp.	*the best*			*the second-best*			*the least accurate*			MST full-scale		
	eigenvalue	*p*	*cp*	eigenvalue	*p*	*cp*	eigenvalue	*p*	*cp*	eigenvalue	*p*	*cp*
1	10.10	25.26	25.26	8.84	22.10	22.10	6.74	16.86	16.86	13.12	32.79	32.79
2	1.95	4.87	30.13	1.80	4.50	26.60	1.63	4.09	20.95	1.73	4.32	37.12
3	1.48	3.71	33.83	1.54	3.84	30.44	1.51	3.77	24.71	1.35	3.36	40.48
4	1.37	3.43	37.26	1.43	3.58	34.02	1.46	3.66	28.37	1.23	3.07	43.55
5	1.29	3.23	40.49	1.33	3.33	37.35	1.43	3.58	31.95	1.09	2.73	46.28
6	1.21	3.04	43.53	1.30	3.25	40.61	1.32	3.29	35.24	1.07	2.68	48.96
7	1.18	2.95	46.47	1.22	3.04	43.65	1.31	3.26	38.51	1.02	2.55	51.51
8	1.12	2.80	49.27	1.20	3.00	46.65	1.25	3.13	41.64	-	-	-
9	1.06	2.65	51.92	1.12	2.79	49.44	1.20	2.99	44.63	-	-	-
10	1.04	2.60	54.53	1.10	2.74	52.18	1.15	2.87	47.49	-	-	-
11	1.02	2.56	57.08	1.02	2.55	54.74	1.14	2.84	50.34	-	-	-
12	-	-	-	1.01	2.52	57.26	1.03	2.58	52.91	-	-	-
13	-	-	-	-	-	-	1.02	2.55	55.46	-	-	-

Note. comp. – number of components extracted; *p* – percent of variance accounted for; *cp* – cumulative percent of variance accounted for.

Relative to the remaining components (whose reliability exceed zero), the first component within each task, and full-scale scores accounted for a substantial proportion of the items' variance. The majority of items within each task have shown high loadings on their principal component. In line with homogeneity parameters, first principal components accounted around 64%, just about ½, 40%, and 30% of the reliable variance of the items in the full-scale, and *the best*, *the second-best*, and *the least accurate* task, respectively.

In order to test the items within three types of tasks in more detail, data were subjected to an IRT-based examination. The basic parameters of the analysis are given in Table 5. Both the indicators of reliability, as well as mean infit and outfit measures have shown to be adequate[8]. Parameters of item and person separation indicated a satisfactory item, but somewhat lower participant differentiation.

Table 5
Reliability, mean infit and outfit and separation indices for items and participants within each task

	the best		*the second-best*		*the least accurate*	
	person	item	person	item	person	item
reliability	0.89	0.98	0.89	0.98	0.86	0.97
mean infit	0.99	0.98	1.00	0.99	1.00	1.00
mean outfit	1.07	1.07	1.02	1.02	1.01	1.01
separation	2.74	7.30	2.79	6.28	2.43	5.63

With the exception of one item in *the best* solution task (infit > 1.30), all the items within all tasks manifested adequate prediction of responses that are close to one's ability level. In terms of the outfit, however, a number of items deviated from the model. The number of unpredicted responses that are far above/below one's ability was the largest for *the best* solution (6 items < 0.70, 10 items > 1.30), lower for *the second-best* (2 items < 0.70, 5 items > 1.30), and the lowest for *the least accurate* solution (2 items > 1.30). However, after detailed examination of the item and person misfits, and the deletion of outliers (mostly low performers who "got lucky" on more difficult items, and a few high performers who made careless mistakes), it was obvious that the items themselves were not at fault for these deviations from the model.

General discussion

This paper presents the process of development and empirical evaluation of the figural reasoning test in the form of matrices with multiple solutions designed for measuring fluid intelligence. Traditional matrices represent one of the best and widely used measures of intelligence (Carroll, 1993; Jensen, 1998; Mackintosh, 1998; Raven, 2000; Spearman, 1946; Vernon & Parry, 1949). Standard highly *G*-saturated tests proved to be good predictors of a wide range of relevant criteria (Gottfredson, 1997; Jensen, 1998; Kuncel et al., 2004; Salgado et al. 2003; Schmidt & Hunter, 1998). However, these tests are often subjected to criticism of not measuring a wide enough range of skills and abilities needed for real-life problem-solving (Ceci, 1990; Gardner, 1993; Sternberg, 1985; Sternberg & Wagner, 1986). One of the prominent criticisms regarding traditional intelligence tests is that they face a person with a problem

8 Item parameters indicate how well a given model predicts the answers to items, while the person parameters indicate the consistency of individual pattern of responses through all items (Embertson & Reise, 2000). Average values for infit and outfit for tests well-set are 1 (Osteen, 2010).

where single one out of limited options available is "the correct" one, making all other options equally wrong (Bujas, 1966; Bujas et al., 1967; Sternberg, 1999).

A satisfactory compromise between opposing perspectives can be found within TVR (Bujas, 1966; Bujas et al., 1967), which in addition to measuring fluid ability, introduced additional requirements which aimed to capture cognitive flexibility expressed through sensitivity to a problem. Resolving some of the issues TVR was facing, the instrument developed within this study aim not to alter the psychometric measures of intellectual abilities, but to complement them by introducing additional requirements that would capture core intellectual abilities more comprehensively. Namely, it can be assumed that by enabling one to deal with a problem from different perspectives, to face and compare alternatives, grade and choose between options of varying degrees of accuracy, the test would presumably represent broader, more flexible and potentially more valid measure of intelligence which is closer to real-life. The rationale for such a presumption can be found in the features of alternative tasks. There is no doubt that the best solution task demonstrates face validity for the assessment of *Gf*, while the other two tasks besides from aiming to measure *Gf*, tend to capture additional processes employed in everyday real-life problem-solving more comprehensively than the standard tests do. In terms of processes involved in solving three types of tasks, the main focus of the best task seems to be the identification and acquisition of the rules that apply to a given matrix, i.e., "correspondence finding" or rule induction (Carpenter et al., 1990), while two alternative tasks are likely to put more load to applying detected rules and coordination between them (see Babcock, 1994) in order to transfer attained inferences to additional complex problems. Therefore, it can be assumed *the second-best* and the *least accurate task* put more load on a variety of executive functions reflected in coordination and monitoring processes (see Babcock, 1994; Carpenter et al., 1990). That being said, alternative tasks of MST should not be seen as measures of fundamentally different processes and abilities than the standard task but as an addition to the salient measure of *Gf*. As such, alternative tasks within the MST could potentially offer more equal coverage of at least three essential executive processes, i.e., updating, shifting, and inhibition (as defined by Miyake & Friedman, 2012; Miyake et al., 2000) that are likely required for many real-life intelligent behaviors and which are not equally assessed by standard intelligence tests (see Friedman et al., 2006). Of course, in order to make definite conclusions on the real-life relevance of alternative tasks evidence on their predictive value over standard matrices and other intelligence tests needs to be demonstrated.

Items of the MST were carefully designed, and went through several stages of evaluation during the process of development, using both qualitative as well as quantitative methods. A large set of items developed has proved to be a fairly good initial point for a completion of the final 40-item instrument. The designed instrument has shown good psychometric properties in all three tasks that are similar to other intelligence tests of this type and length. Overall, reliability indicators for all three tasks lie in the range of reliabilities of well-established matrices

like Raven's Progressive Matrices (Domino & Domino, 2006; Raven, Raven, & Court, 1998; Raven, Raven, & Court, 2000) and Wechsler's matrices (Sattler & Ryan, 2009; Wechsler, 2008), while the reliability measures of individual tasks are equal to or slightly higher than the full-scale reliability of TVR (Dragićević & Momirović, 1969). Furthermore, results clearly indicate the existence of a certain amount of shared variance between the items within each task, which derives from their core similarities, as well as common abilities engaged in solving them. However, within each task, a few items have exhibited a relatively high proportion of task-specific variance. This fact is most likely the result of the large diversity of rules applied to items and contents of individual items.

High homogeneity and reliability of full-scale instrument provide grounds for the calculation and usage of the integral test score only. Of course if one is interested in characteristics and specificities of individual tasks calculation of three separate scores is a possibility. In that case, one has to bear in mind that participants' answers within items are not independent of each other so the results should be interpreted accordingly.

On the other hand, discriminative power has shown to be relatively poor. However, it should be noted that discriminative power, thus all other psychometric properties of the test were strongly influenced by the sample as well as the full-test's characteristics. Firstly, the distortion of the distribution of scores in some of the tasks is the most certainly induced by characteristics of the sample (university educated individuals) leading to a restriction in variance, and the reduction in all parameters obtained. Thus, discriminative power would certainly be much higher in a heterogeneous sample selected from general population where *the best* solution task would probably be normally distributed, while *the second-best* and *the least accurate* task would be somewhat more difficult. Secondly, it can be assumed that, to some extent, participants' errors on the more difficult items (the second half of the test) can be attributed to their inability to maintain attention for a prolonged period of time needed to complete the test. Additionally, this could be the reason for somewhat lower, but still fair psychometric properties of *the least accurate* task in comparison to the other two. It's safe to assume that participants were probably solving this task the last within each item, and it's possible that some of the errors were primarily the result of the participants failing to maintain their focus.

It should be noted that the instrument designed within this study, besides conceptual, offers at least two practical novelties that cannot be found among other intelligence tests. Firstly, out of the items developed in this study a researcher could easily construct three separate tests each of which could stand for itself, requiring only one type of answer. In that way, one could focus on the different aspects and properties of problem-solving – the ability to detect *the best* solution, ability to detect only *the second-best* solution, or measuring ability to detect *the least accurate* solution. Bearing in mind that results suggest that all three tasks measure same core abilities but different aspects of problem-solving this asset of the instrument can be further exploited in different ways. Secondly, one can arrange the instrument in a way so as to have the subjects face different

requirements in different items (for example, for item *X* a participant could be asked to find only *the best* solution, for item *Y* only *the second-best*, for item *Z* just *the least accurate* one, etc.). Finally, having in mind large reliability indices of aggregated full-scale score one could construct several shorter parallel tests in the same form as presented within this paper and to be able to economically assess subject's intelligence in repeated testing when the sufficient time period between successive testing is an issue.

In closing, it's important to note that the evaluation of the 40-item version of the test needs to be carried out on an independent sample of participants since it is possible that results of the final version were impacted by the full-test's content, its length, sequence effects, etc. Additional evaluation of the final instrument should be carried out on an independent gender-balanced sample from the general population, thus validating results obtained in this study on a sample with more diverse levels of abilities.

In the second part of the study, we explore the validity of the developed instrument and provide data on construct and predictive validity of the MST (see Živanović et al., 2018, this issue of Psihologija).

Conclusion

Matrices in the form of the MST are offering a reliable, more flexible, and a potentially broader measure of intelligence than the standard form of the matrices. Although in comparison to traditional intelligence tests, the test-situation remains the same, nevertheless the alternative tasks offer one a possibility to express abilities in a much broader and flexible context, which can be, by means of the number of "correct" options available and the abilities needed to detect solutions of varying degrees of accuracy, considered similar to the requirements of real-life problem-solving.

References

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

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219–254. doi:10.1037/0033–2909.121.2.219

Babcock, R. L. (1994). Analysis of the adult age differences on the Raven's Advanced Progressive Matrices Test. *Psychology and Aging, 9*, 303–314. doi:10.1037/0882–7974.9.2.303

Bujas, Z. (1966). *Modifikacija Ravenovih Progresivnih Matrica* [Modification of the Raven's progressive matrices]. Zagreb: Odsjek za psihologiju.

Bujas, Z., Bartolović, B., & Vodanović, M. (1967). *Test višestrukih rješenja* [Multiple solution test]. Zagreb: Odsjek za psihologiju.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In P. D. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 122−130). New York: Guilford.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In P. D. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (2nd edition) (pp. 69−76). New York: Guilford.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. New York: North-Holland.

Ceci, S. J. (1990). *On intelligence... more or less: A bio-ecological treatise on intellectual development*. Englewood Cliffs, NJ: Prentice Hall.

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence, 40*(3), 278–295. doi:10.1016/j.intell.2012.02.010

Colom, R. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*(3), 277–296. doi:10.1016/j.intell.2003.12.002

Colom, R., Abad, F. J., Rebollo, I., & Shih, P. C. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence, 33*(6), 623–642. doi:10.1016/j.intell.2005.05.006

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. . (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183. doi:10.1016/S0160–2896(01)00096–4

Denis, P. L., & Gilbert, F. (2012). The effect of time constraints and personality facets on general cognitive ability (GCA) assessment. *Personality and Individual Differences*, *52*(4), 541–545. doi:10.1016/j.paid.2011.11.024

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*, 135–168. doi:10.1146/annurev-psych-113011–143750.Executive

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction (2nd Edition).* Cambridge University Press.

Dragićević, Č., & Momirović, K. (1969). *Standardizacije NSI i RBM na teritoriji SR Srbije* [Standardization of NSI and RBM in Serbia]. Beograd: Institut za psihologiju.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331. doi:10.1037/0096–3445.128.3.309

Friedman, N. P., Miyake, A., Corley, R., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are, related to intelligence. *Psychological science, 17*(2), 172–179. doi:10.1111/j.1467–9280.2006.01681.x

Gardner, H. (1993). *Frames of mind: The Theory of multiple intelligences*. New York: Basic Books.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence, 52*, 71–79. doi:10.1016/j.intell.2015.07.006

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132. doi:10.1016/S0160–2896(97)90014–3

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179−203. doi:10.1016/0160–2896(84)90008–4

Gustafsson, J. E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 4* (pp. 35–71). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Gustafsson, J. E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetriou, & A. Efklides (Eds.), *Intelligence, mind and reasoning: Structure and development.* Elsevier. doi:10.1016/S0166–4115(08)62752–1

Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73−95). London: Lawrence Erlbaum Associates, Publishers.

Horn, J. L. (1979). The rise and fall of human abilities. *Journal of Research and Development in Education, 12*(2), 59–78.

Horn, J. (1988). Thinking about human abilities. In J. R. Nesselroade, & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 645–685). New York: Plenum Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253−270. doi:10.1037/h0023816

Jensen, A. R. (1982). *Reaction time and psychometric g*. In H. J. Eysenck (Ed.), A model for intelligence (pp. 93–132). New York: Springer.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, Connecticut: Praeger Publishers.

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*(1), 66–71. doi:10.1037/0033–2909.131.1.66

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–217. doi:10.1037/0096–3445.133.2.189

Knežević, G., & Momirović, K. (1996). RTT9G i RTT10G: dva programa za analizu metrijskih karakteristika kompozitnih mernih instrumenata [RTT9G and RTT10G: Two programs for the analysis of metric properties of composite measuring instruments]. In: P. Kostić, *Merenje u psihologiji, 2* (pp. 35–56). Belgrade: Institute for Criminological and Sociological Research.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148–161. doi:10.1037/0022–3514.86.1.148

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389–433. doi:10.1016/S0160–2896(05)80012–1

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29–37. doi:10.1111/j.1745–3992.2007.00106.x

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to "simple" short-term storage? Intelligence, *39*(6), 473–480. doi:10.1016/j.intell.2011.09.001

Mackintosh, N. J. (1998). *IQ and Human Intelligence*. UK: Oxford University Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. doi:10.1016/j.intell.2008.08.004

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*(7), 651–675. doi:10.1002/pits.20497

Miyake, A., & Friedman, N. P. (2012). The nature and organisation of individual differences in executive functions : four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. doi:10.1177/0963721411429458.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. doi:10.1006/cogp.1999.0734

Oberauer, K., Süβ, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence, 36*(6), 641–652. doi:10.1016/j.intell.2008.01.007

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, *1*(2), 66–82. doi:10.5243/jsswr.2010.6

Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology, 16*(2), 97–104. doi:10.1111/j.2044–8341.1936.tb00690.x

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. London: H.K. Lewis.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive psychology*, *41*(1), 1–48. doi:10.1006/cogp.1999.0735

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

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 4. Advanced progressive matrices*. Oxford: Oxford Psychologist Press.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Raven manual: Section 3. Standard progressive matrices*. Oxford: Oxford Psychologist Press.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*(6), 1068–1081. doi:10.1037/0021–9010.88.6.1068

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS–IV*. San Diego, CA: Jerome M. Sattler, Publisher.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. doi:10.1037/0033–2909.124.2.262

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

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 13*, 201–293. doi:10.2307/1412107

Spearman, C. E. (1927). *The abilities of man*. London: Macmillan.

Spearman, C. E. (1946). Theory of the general factor. *British Journal of Psychology, 36*, 117–131. doi:10.1111/j.2044–8295.1946.tb01114.x

Sternberg, R. J. (1985). *Beyond IQ: A Triarchic theory of human intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology, 3*, 292–316. doi:10.1037/1089–2680.3.4.292

Sternberg, R. J., & Wagner, R. K. (1986). *Practical Intelligence: Nature and Origins of competence in the everyday world*. New York: Cambridge University Press.

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2*(1), 1–5. doi:10.1111/1467–8721.ep10770441

Süß, H. M., & Beauducel, A. (2015). Modeling the construct validity of the Berlin Intelligence Structure Model. *Estudos de Psicologia* (Campinas), *32*(1), 13–25. doi:10.1590/0103–166X2015000100002

Vernon, P. E., & Parry, J. B. (1949). *Personnel selection in the British Forces*. London: University of London Press.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Pearson Assessment.

Živanović, M., Bjekić, J., & Opačić, G. (2018). Multiple solutions test Part II: Evidence on construct and predictive validity. *Psihologija.* https://doi.org/10.2298/PSI170205004Z

RECEIVED 31.10.2016.
REVISION RECEIVED 10.08.2017.
ACCEPTED 14.09.2017.

Test višestrukih rešenja
Deo I: Razvoj i psihometrijska evaluacija

Marko Živanović[1], Jovana Bjekić[2] i Goran Opačić[1]

[1]*Odeljenje za psihologiju, Filozofski fakultet, Univerzitet u Beogradu, Srbija*
[2]*Institut za medicinska istraživanja, Univerzitet u Beogradu, Srbija*

Kako se ljudi van testovnog konteksta retko nalaze u situacijama gde su suočeni sa ograničenim opcijama i samo jednim tačnim odgovorom, pri čemu su svi ostali odgovori podjednako pogrešni, izmena tradicionalnih testova inteligencije (u smislu povećanja fleksibilnosti) može potencijalno da obezbedi obuhvatniju i validniju meru inteligencije. Stoga, cilj ove studije je razvoj i psihometrijska evaluacija testa figuralnog rezonovanja u formi matrica sa višestrukim rešenjima. Za razliku od konvencionalnih testova inteligencije, u ovom testu ispitanici su suočeni sa više od jednog zadatka, tj. oni treba da otkriju: 1) *najbolje rešenje* – figuru koja najbolje kompletira datu matricu; 2) *drugo najbolje rešenje* – figuru koja bi najbolje upotpunila matricu ukoliko ne bi bilo prvog, najboljeg odgovora; 3) *najmanje tačnu opciju* – figuru koja upotpunjava datu matricu na najmanje tačan način. U procesu razvoja testa, konstruisan je početni skup od 80 stavki i zadat uzorku od 41 ispitanika, sa ciljem sticanja uvida u kvalitet i potrebu za prilagođavanjem početnog skupa stavki. Psihometrijske karakteristike instrumenta koji se sastoji od 74 stavke sa tri tipa zadataka proverene su na uzorku od 263 ispitanika, nakon čega je predložena kratka verzija instrumenta. Sva tri zadatka unutar testa, kao i test u celini pokazali su dobra interna psihometrijska svojstva ($\alpha_{\text{najbolje rešenje}} = .92$, $\alpha_{\text{drugo najbolje rešenje}} = .90$, $\alpha_{\text{najmanje tačno rešenje}} = .87$; $\alpha_{\text{ukupni skor}} = .95$) nudeći mogućnost pouzdanog i obuhvatnijeg merenja inteligencije.

Ključne reči: Test višestrukih rešenja (MST), figuralno rezonovanje, fluidna sposobnost (*Gf*), razvoj testa, psihometrijska svojstva.