Psychological Test and Assessment Modeling, Volume 64, 2022 (1), 112-126

Alea iacta est - Development and evaluation of a manipulation approach of the item difficulty in the distractor-based format of the cube construction test

Julie Levacher[1,2], *Marco Koch*[2], *Maximilian Bungart*[2], *Frank M. Spinath*[2] & *Nicolas Becker*[3]

Abstract

Spatial reasoning is a component of intelligence with very high loadings on the g factor and which has high practical importance for personnel selection. To this extent, much research has been conducted regarding the measurement of spatial reasoning. Oftentimes, tests require participants to construct a mental representation of an object and to manipulate that object in order to solve an item. Past endeavours to develop distractor-based tests have resulted in rather narrow distributions of item difficulties, reducing their usefulness for selection procedures. Thus, in the current study, one test of spatial reasoning has been administered to a sample of $N = 116$ psychology students with the aim of developing a manipulation approach to provide a broad range of item difficulties. The resulting items were of average difficulty ($p_i = .55$) with a wide range ($.18 \leq p_i \leq .92$). Compared to previous research, item parameters are in a range that is more useful for selection procedures. Albeit no predictor for item difficulty could be identified, an exploratory analysis suggested that this might be caused by non-overlapping items between the original cube construction task and the newly developed distractor based task. Nonetheless, this study provided a new approach to the manipulation of item difficulties in spatial reasoning tasks which can benefit practitioners when conducting personnel selection and researchers when developing new diagnostic instruments.

Keywords: spatial reasoning, mental rotation, test-development, distractor-based tests, item parameter manipulation

[1] *Correspondence concerning this article should be addressed to:* Julie Levacher, Campus A1.3, 66123 Saarbrücken, Germany, julie.levacher@uni-saarland.de

[2] Department of Individual Differences & Psychodiagnostics, Saarland University, Saarbrucken, Germany

[3] Department of Individual Differences & Psychodiagnostics, Greifswald University, Greifswald, Germany

Introduction

The goal of this study is to develop and evaluate a manipulation approach for a distractor-based intelligence test and focussing on the challenge of establishing a broad range of item difficulties. In distractor-based cube rotation tests the reference cubes are presented together with some filled cross sections as response options. The task is to decide which of the cross sections, represents the reference cube. Those distractor-based performance tests are associated with some drawbacks such as response-elimination and guessing (Carpenter et al., 1990; White & Zammarelli, 1981) but also have the advantage that they can be used better in group settings because of their presentation as paper-pencil test. For computer-based testing, Thissen and colleagues (2018) developed a distractor-free test of spatial reasoning. However, due to the very high difficulty of the test, a further study attempted to develop items of a broader difficulty range (Thissen et al., 2019). Simultaneously, they also tested a distractor-based variant of their test. While they reported an association between item difficulty and construction principles for the distractor-free variant, this information was not reported for the distractor-based version. Furthermore, in the distractor-based version only a very narrow range of item difficulties was found. The present study therefore attempts to generate distractor-based items in such a way that there is an equally broad range of difficulties as in the distractor-free version and that those difficulties can be predicted from the construction principles.

Spatial Ability

Spatial Ability refers to the perception and processing of visual images and includes, for example, the ability to estimate length and to recognise spatial relationships (spatial relations), but also visual imagination (spatial visualisation), the ability to imagine the appearance of a stimulus from another perspective (spatial orientation) as well as an efficient spatial search. The three close factors spatial relations, spatial visualisation, and spatial orientation have already been found in two large meta-analyses on spatial reasoning (Carroll, 1993; Lohman, 1979). Furthermore, spatial visualisation tasks show a strong association with the *g* factor of intelligence (Horn, 1968; Lohman, 1996; Lohman, 1988). Since *g* is also one of the strongest predictor of school-grades(Roth et al., 2015), it could be expected that tests of spatial ability should also be associated with *g*. Tasks with a high *g*-saturation are primarily matrix tasks or spatial imagination tasks (Formann & Piswanger, 1979; Hossiep et al., 1999; Schmidtke & Raven, 1978). But the cube construction task can also be classified accordingly (Thissen et al., 2018). The connection of *g* and spatial abilities is potentially influenced by an efficient (logical) reasoning (Grüßing, 2002).

Cube Rotation Items

Cube rotation tasks are mainly used to assess spatial ability (Amthauer et al., 2001; Gittler, 1990; Jäger et al., 1997). In classical items, such as those of the "Three-Dimensional Cube Test" by Gittler (1990), a reference cube is presented together with several response cubes. The task is to identify the response cube that is identical to the reference cube by mental rotation (see **Figure 1**). However, only one of the given cubes is correct at a time. All other cubes presented are only distractors.

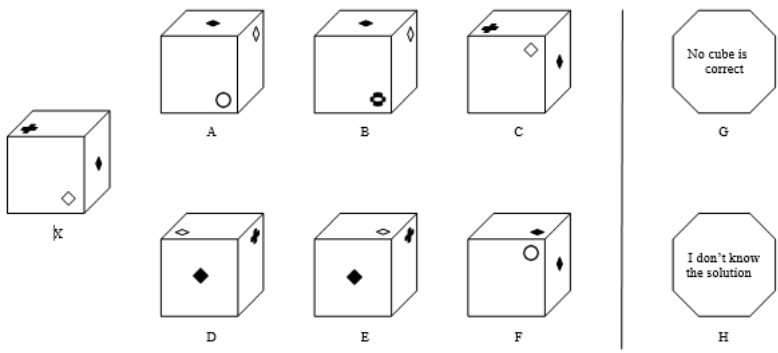

Figure 1. Classic answer format in cube rotation tasks. The correct answer is D. (according to Thissen et al., 2016)

Problems Concerning the Manipulation of Item Difficulties

The use of distractors gives rise to several problems, which can have a particularly negative effect on the validity of the test. An unavoidable problem of distractor-based task formats is the existence of a guessing probability. This means that every respondent can solve the tasks by simply guessing, almost regardless of his or her cognitive ability. The guessing probability is 1/k, where k is the number of response options. Therefore, correct scores derived from guessing dilute the validity of the test result. This reduces the construct validity of the test, since guessing is not an adequate indicator for the construct to be measured. Simultaneously, it also reduces the difficulty of the respective task to a certain extent.

There are two possible solution strategies for responding to distractor-based tasks (Arendasy & Sommer, 2013; Becker et al., 2016; Vigneau et al., 2006). Response elimination describes the strategy to compare single elements of two objects to exclude wrong answers. In relation to the cube rotation task, it means that a respondent

successively compares the response cubes with the reference cube to identify mismatches. In this way, wrong answer options can be systematically excluded until either only one solution remains, or a guess can be made due to only a small number of remaining answer options. The second response strategy is called constructive matching. It is defined by actively constructing the possible answer, that can be subsequently compared with the presented answer options. In this way, the one correct solution can be selected. In relation to the cube rotation task, this means that first a mental model must be generated from the reference cube. Sequentially this model can be rotated mentally and can be compared with the response cubes. Embretson (1994) described this solution process similarly. After encoding the item stem and the response options, an anchor (two adjacent sides) is searched and matched to the stem. Then necessary rotations are carried out in order to fold the two-dimensional cube from the response options. Finally it can be confirmed whether there are mismatches or whether the cubes are identical. This is the desirable strategy for solving the tasks as it does not confound the mental processes used for solving mental rotation with processes used for (two-dimensional) visual comparison.

Another problem for the development of mental rotation tasks is the lack of strategies to manipulate item difficulty. For example, the difficulties of the items in the cube rotation tasks without a time limit are quite homogeneous and moderate (M_p = .44, SD_p = .10; Gittler, 1990). One reason for this can be seen in the uniform construction method. In contrast to figural matrices, no construction rules can be combined here, so that changing the symbols is the only way to manipulate the difficulty. The absence of items with high and low difficulty leads to the problem of reduced differentiability in groups with high and low ability (Mittring & Rost, 2008). The result is therefore a reduction in convergent validity with respect to other ability tests that have more variance in item difficulty. Also, a limitation on the given completion time of the entire test does not lead to a variation in difficulty. It is possible that some items can no longer be completed, but not because they are more difficult, but simply because of time limitations (Wilhelm & Schulze, 2002). This means that the results are influenced by a mental speed component. In general, it was shown that this speed-and-power-problem in a challenge in psychological assessment (Glück & Fabrizii, 2010).

Solving the problems of distractor-based task formats

The problems regarding the manipulation of item difficulties can be solved in several ways. Firstly, the corresponding response format can be adapted. This solution was chosen by Thissen and colleagues (2018). For this purpose, the authors replaced the distractor-based response format with 30 individual symbols (see **Figure 2**). This meant that the test persons could not select from various answer options but had to construct the corresponding answer themselves.

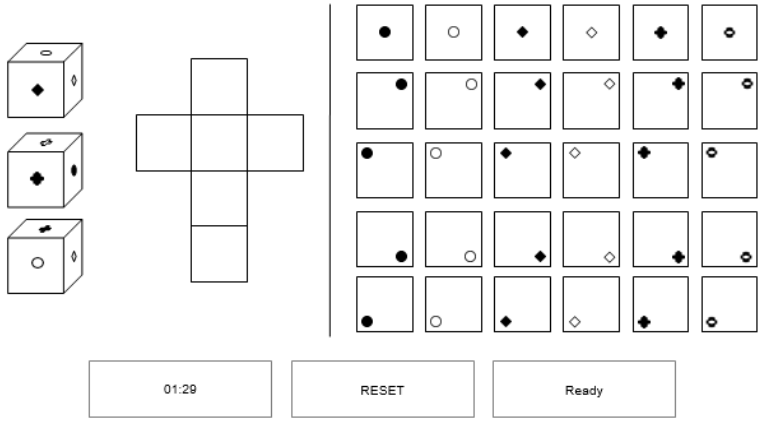

Figure 2. The distractor-free cube rotation task (according to Thissen et al., 2016)

This is therefore a distractor-free procedure. The reference cube is also presented from three different perspectives to ensure an unambiguous solution. The subject's task is to relate the three perspectives of the cube to each other (spatial relations) and to mentally unfold and rotate them at the same time (spatial visualisation). Overall, the solution process corresponds to the constructive matching strategy. Only after a mental model of the cube has been created is it possible to reconstruct the answer with the help of a symbol pool within an empty cross section. For this task, 90 seconds are available.

But even if the distractor-based format is to be retained, there are ways to vary the difficulties. A first approach related to cube rotation tasks was also pursued by Thissen and colleagues (2018). For this, the identical item stems as in the construction-based task are chosen. Instead of an empty cross section as an answer field in combination with a symbol pool, however, six answer options are depicted here. These represent six fully expanded cube fields. In addition, following Gitter (1990) the categories "No cube is correct" and "I don't know the solution" are offered to minimise the guessing probability (see **Figure 3**).

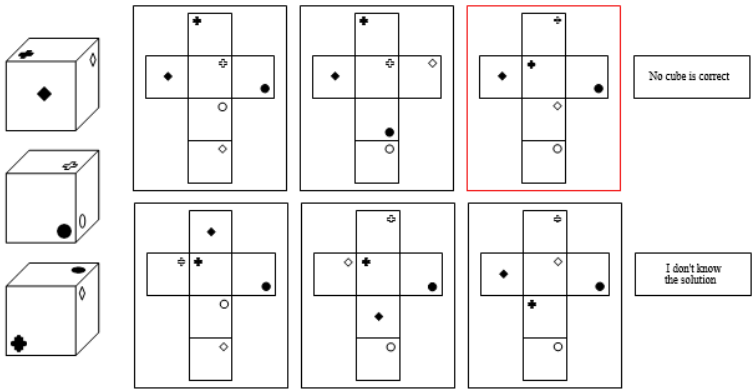

Figure 3. Distractor-based format of the cube construction test with correct solution highlighted on top right (according to Thissen et al., 2018)

Both test formats showed comparable internal consistency (distractor-based $\alpha = .81$ vs. distractor-free $\alpha = .83$) and selectivity (distractor-based $M(r_{it}) = .40$ vs. distractor-free $M(r_{it}) = .36$). Mean item difficulties differed significantly from each other (distractor-free $M(p_i) = .27$ vs. distractor-based $M(p_i) = .40$) and the difficulty range was also wider in the distractor-free format ($.02 \leq p_i \leq .95$) than in the distractor-based one ($.37 \leq p_i \leq .63$). Consequently, subjects solved significantly more items in the distractor-based version than in the distractor-free version. The current study continues from this point and attempts to manipulate the difficulties of the distractor-based version so that a greater range of difficulty should result.

Goals and Research Questions of the Study

The focus of this study is the development and evaluation of a manipulation approach of item difficulty within the distractor-based response format of the cube construction task of Thissen and colleagues (2019). In general, one can differentiate between perceptual distractors (i.e., they visually resemble the correct response option) and conceptual distractors (i.e., they employ the same rules as the correct solution but contain intentional deviations; Becker et al., 2016). The chosen approach for manipulation is similar to the one already used in the distractor-free format, i.e., the difficulty should be varied by the number of default symbols. Contrary to the hypotheses of Thissen and colleagues (2019) it is assumed that the fewer default symbols are given, the easier the task is to be solved, as less rotation is required. This represents a mixture of perceptual and conceptual distractors, because an increasing number of default symbols simultaneously increases the resemblance between distractors and target and also

leads to more mental rotation being necessary. With the framework of Embretson (1994) it can be argued that less symbols mean it is easier to identify an anchor in the response options and only few rotations are needed for verifying two cubes' identity. The focus on the number of default symbols allows for a very systematic variation of item features and holds the potential to predict item difficulty compared to regular, purely perceptual distractors. Therefore, we focussed on the following research questions:

Question 1: Is the test procedure resulting from the difficulty manipulation reliable? More specifically, does the internal consistency exceed a minimum value of Cronbach's $\alpha = .80$?

Question 2: Is there a significant negative correlation between the item difficulty and the number of default symbols?

Question 3: Is it possible to increase the range of item difficulty within the distractor-based cube construction task?

Question 4: Does the test measure a single underlying factor?

Methods

Sample

The sample consisted of 116 psychology students (81% female, 19% male) from Saarland University with a mean age of 22.46 years ($SD = 3.01$, $18 \leq \text{age} \leq 35$). No participants were excluded from further analyses. In addition to age and gender, the grade point average (GPA) was also recorded as an external criterion variable ($M = 1.76$; $SD = 0.51$, $1.0 \leq \text{GPA} \leq 3.3$). For participating in the study, the participants received partial course credit.

Manipulation of the item difficulty

The 23 item stems, the response options, and the arrangement of the distractor-based format of Thissen and colleagues (2019) were used as a basis. As a manipulation approach, the number of given (default) symbols on the unfolded cubes was also varied here. The difference, however, is the contrary assumption that tasks with fewer default symbols are easier to solve than tasks with increasing numbers. The reason for this assumption can be seen in the mental rotation. Therefore, the first difficulty level (two and three default symbols) should be easier to solve than the last difficulty level (five and six default symbols). Tasks with four default symbols should therefore be in the middle difficulty range. Altogether five different levels of difficulty were varied. Each

of these levels contained between four and five items. All items including the solution as well as their level of difficulty can be found in the ESM[1].

For creating the answer options, symbols were selected sequentially and deleted from all unfolded cubes in the response options. This was done until the number of symbols of the corresponding difficulty level was reached. As in the study by Thissen and colleagues (2019), the selection of the deleted symbols was also done randomly. Since no uniform deletion rule could be defined, care was taken when deleting the symbols to avoid duplicate solutions within the response options.

However, in the example task and item 1 (each with two default symbols), this procedure led to the fact that the symbols of the distractors had to be placed differently to avoid multiple solutions. How this was implemented in the example task is shown in the ESM[1]. For the remaining items of the first level of difficulty a change of position was not necessary, because there were some symbols in a corner position and no double solutions were produced by deleting the symbols. Beyond the changes mentioned above, all items were kept in their original form. The solution was finally randomly assigned to one of the six answer alternatives (a to f).

Procedure

The survey was conducted in a laboratory setting in small groups. The test was allowed to take place on the subjects' own laptops and began by collecting some demographic information. Afterwards the instruction of the test followed immediately on the screen. To ensure understanding, the test included an instruction phase. Additionally, a test administrator was present. Throughout the test, the number of default symbols increased steadily. Processing time was set to 180 seconds per item. If the person answered within the time window, the remaining time was also recorded. If they did not answer within the given time window, the task was considered as not solved. Each task was followed by a skippable pause of 5 seconds before the next item was automatically displayed. The test thus had a maximum duration of about 70 minutes.

Statistical Procedure

The internal consistency of the distractor-based Cube Construction Task is calculated as Cronbachs's Alpha. In addition, the selectivity of the items was calculated by forming the correlation of each item with the rest of the scale (part-whole correlation).

To check whether the approach of difficulty manipulation was successful, a Pearson correlation between the solution probabilities and the difficulty levels (i.e., number of default symbols) was calculated. Regarding item difficulty, the solution probability or difficulty index Pi was calculated for each item, which represents the percentage of correct answers for item i in a sample of size n (Pospeschill, 2010). To check whether

the test was based on a single factor structure, a confirmatory factor analysis was calculated. All statistical analyses were performed with raw data by using SPSS. The used standard significance level was 5% ($p < .05$) and confidence interval level was set to 95 %.

Results

Reliability

The test demonstrated high internal consistency ($\alpha = .82$) and an average part-whole-corrected item total correlation coefficient of $M(r_{ph}) = .37$, $SD = .10$, $10 \leq r_{it} \leq .56$. Values for each item can be found in **Table 1**. The reliability of the test can be regarded as given.

Test and item difficulty

Regarding the number of correct responses, the mean test score was 12.61 ($SD = 4.96$), with a minimum of 0 and a maximum of the 23 possible correct responses. Therefore, the sample consisted of people with average ability with a few participants who were unable to solve any items and a few participants who could solve all items. Concerning the possible processing time of 70 minutes it could be shown that the participants needed on average 39.86 minutes ($SD = 9.75$). Additionally, no significant correlation between the test score and the processing time could be found ($r = .13$, $p = .168$). Likewise, the correlation of the test value with the GPA ($r = -.066$, $p = .479$) and age ($r = -.010$, $p = .919$) was not significant.

The mean item difficulty was $M(p_i) = .55$ with a standard deviation of $SD(p_i) = .16$. The range of item difficulties was $.18 \leq p_i \leq .92$. The item difficulties are listed in **Table 1** together with the number of default symbols and the respective item selectivity.

Table 1. Item characteristics of the distractor-based cube rotation task

Item	*DS*	Σ(*R*)	p_i	r_{it}
1	2	107	.92	**.29**
2	2	76	.66	**.32**
3	2	54	.47	**.30**
4	2	77	.66	.43
5	3	82	.71	**.37**
6	3	41	.35	**.29**
7	3	78	.67	**.34**
8	3	72	.62	.51
9	3	41	.35	**.34**
10	4	51	.44	.50
11	4	21	.18	**.10**
12	4	41	.35	.41
13	4	52	.45	.50
14	4	68	.59	**.39**
15	5	78	.67	.43
16	5	66	.57	**.34**
17	5	53	.46	**.25**
18	5	62	.53	.42
19	5	66	.57	.44
20	6	62	.53	.56
21	6	62	.53	**.38**
22	6	72	.62	**.26**
23	6	81	.70	**.36**
M	4	63.13	.55	.37
SD	1.38	17.88	.16	.10

Notes. *DS* = default symbols; Σ(*R*) = Total of the correct solutions; p_i = item difficulties; r_{it} = item selectivity (marked when below .40)

The correlation between item difficulty and the number of default symbols was not significant (r = -.144, p = .605). From this it can be concluded that the difficulty manipulation was not successful according to the hypothesis.

Factor validity

In addition to the high internal consistency, it can also be assumed that the result of a factor analysis could indicate a unidimensional latent construct. With reference to research question 4, a confirmatory factor analysis was calculated. This was performed using the estimation algorithm WLMSV. The results of the analysis are shown in Table 2 and confirm the single factor solution.

Table 2. Fit indices of the confirmatory factor analysis

χ^2	*df*	*P*	*CFI*	*RMSEA*	*SRMR*
250.6	230	.168	.963	.028	.124

Notes. χ^2 = Chi-square; df = degrees of freedom; p = significance level; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

The Chi-square test was not significant, which means that the empirical model does not deviate from the expected one, χ^2 (230) = 250.6, p = .168. It is also important that the ratio of χ^2/df is as low as possible, in this case 1.09. This does not exceed a proposed cut-off value of 3 (Wheaton et al., 1977). According to the conventions of Carlson and Mulaik (1993) as well as Browne and Cudeck (1993), the fit indices are also predominantly in a good range.

Discussion

The aim of the present study was to provide means by which a spatial reasoning task with a broad and predictable range of item difficulties can be generated. To this extent, new distractors have been developed for the 23 original items (Thissen et al., 2019) based on the assumption that providing more default symbols in the response options makes the tasks more challenging as test takers need to carry out more complex mental rotations.

With regard to the first research question, the results suggest that the test conducted in this study is very reliable (α = .82) which ties in well with the results from Thissen and colleagues (2019; α = .81). The average part-whole correlation coefficient ($M_{(rit)}$ = .37) is also comparable to the original results ($M_{(rit)}$ = .37). Since the development procedure was largely a replication of the original study these results therefore support the validity of the distractor generation.

However, the original study reported a rather narrow range of item difficulties ($.37 \leq p_i \leq .63$; Thissen et al., 2019) casting doubt to the usefulness for differentiation in low-ability or high-ability samples. Contrary to those results, the items from the present study with newly developed distractors consisted of items with a broader range of item difficulties ($.18 \leq p_i \leq .92$). This enables the test to not only differentiate between people of average ability but also between people on both ends of the ability distribution. This is especially important if the test is to be used for selection procedures (i.e., student admission tests, personnel selection) as applicants might self-select depending on their own ability.

The present research also tried to predict the item difficulty with the number of default symbols used in the response options. However, there was no evidence for the hypothesized linear trend. Therefore, these findings can be understood as a contrast to the results of Thissen and colleagues (2019) who reported a strong linear association for the distractor-free version ($r = .71$). This result can be explained by a technical difference. While in the distractor-free version one to five default symbols were used, the distractor-based version used two to six default symbols. This is a necessity as six default symbols in the distractor-free version and one default symbol in the distractor-based-version inherently imply the correct answer (or an unsolvable item). An exploratory analysis of the association of item difficulties with default symbols for only overlapping items (i.e., two to five default symbols) results in a moderate negative correlation coefficient ($r = -.36$) which is in line with the hypothesis. More default symbols require more complex mental rotation in order decide whether a response option fits the item stem. However, presenting five default symbols in the response options may render mental rotation obsolete because every item can be solved by visually inspecting the item stem and merely comparing adjacent sides. If these items were to be excluded from the analysis, the correlation coefficient would rise considerably ($r = -.52$). Contrary, in the distractor-free format, items cannot be solved by merely detecting differences between the item stem and response options. Instead, participants are required to insert all required symbols by themselves. Correctly entering five symbols (in the case of one default symbol) requires more mental operations than entering one symbol (in the case of five default symbols).

Disregarding whether items with five default symbols are appropriate for measuring spatial reasoning, the results of a one-factor confirmatory factor analysis suggest, that the task measures one single factor, corroborating the finding of Thissen and colleagues (2019).

With regard to the tests criterion validity, tests of spatial reasoning have been repeatedly shown to be one of the best indicators for *g* (Deary et al., 2010) and *g* has been shown to be a strong predictor of academic achievement (Roth et al., 2015). Surprisingly, in the present study no association was found between the spatial reasoning test score and academic achievement measured as the GPA. Thissen and colleagues (2019) did not provide any estimates for criterion validity, therefore no comparison is possible. It must be noted that psychology students are an extremely selective sample with restricted range in intelligence and GPA reducing the chances of finding an

association. However, as the main purpose of this study was the development of a broad difficulty range and to assess how this difficulty can be predicted from item features, a compliant sample in a controlled setting was preferable to a more heterogeneous sample.

One further limitation of the present study was the fact that the distractor-free task was not used simultaneously as this would have made it possible to inspect whether the response format changes the latent construct measured by the test. While Thissen and colleagues (2019) administered both versions, they did not report a measurement invariance model. Therefore, this is an important question for future research. Furthermore, although our exploratory analyses suggest that an extreme number of default symbols (i.e., six for the distractor-based version or one for the distractor-free version) might motivate the participants to employ other response strategies than constructive matching, there is no conclusive evidence, yet. Future studies might employ think-aloud designs to gain further insights into the processes underlying the cube construction task. Moreover, studies using bigger samples might estimate IRT models in order to gain further insights into the association of item parameters with item design features.

To summarize, the present study suggests that a distractor-based cube rotation task can be developed and administered with an item difficulty distribution that is comparable to the distractor-free version. This is an important contribution because the distractor-free version can be administered on a computer exclusively, making it an unfeasible endeavor for large group-test settings. Furthermore, these results suggest that, despite the inherent guessing rate, the distractor-based cube rotation task can differentiate between people with very high (or very low) spatial reasoning ability. Although further research is needed to clarify whether items with five and six default symbols should be excluded to guarantee a more homogeneous measurement of spatial abilities, the present study provides a test for spatial abilities that can be administered in more diverse settings than the distractor-free version.

[1]ESM see https://osf.io/6gxw8/

References

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). Intelligenz-Struktur-Test 2000 R. Göttingen: Hogrefe, 2.

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

Becker, N., Schmitz, F., Falk, A., Feldbrügge, J., Recktenwald, D., Wilhelm, O., Preckel, F., & Spinath, F. (2016). Preventing Response Elimination Strategies Improves the Convergent Validity of Figural Matrices. *Journal of Intelligence*, *4*(1), 2. https://doi.org/10.3390/jintelligence4010002

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Carlson, M., & Mulaik, S. A. (1993). Trait Ratings from Descriptions of Behavior As Mediated by Components of Meaning. *Multivariate Behavioral Research*, *28*(1), 111–159. https://doi.org/10.1207/s15327906mbr2801_7

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

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

Cohen, J. B. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201–211. https://doi.org/10.1038/nrn2793

Embretson, S. (1994). Applications of Cognitive Design Systems to Test Development. In C. R. Reynolds (Ed.), *Cognitive Assessment (pp. 107–135). Springer US*. https://doi.org/10.1007/978-1-4757-9730-5_6

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT)*. Beltz Test.

Gittler, G. (1990). *Dreidimensionaler Würfeltest: 3 DW;* [ein rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens]. Beltz.

Glück, J., & Fabrizii, C. (2010). Gender differences in the mental rotations test are partly explained by response format. *Journal of Individual Differences*, 31(2), 106–109. https://doi.org/10.1027/1614-0001/a000019

Grüßing, M. (2002). Wieviel Raumvorstellung braucht man für Raumvorstellungsaufgaben?. *Zentralblatt für Didaktik der Mathematik*, 34(2), 37-45.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review*, *75*(3), 242–259. https://doi.org/10.1037/h0025662

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). Berliner Intelligenzstruktur-Test: BIS-Test. Hogrefe.

Lohman, D. F. (1996). Spatial Ability and G. *Human Abilities: Their Nature and Measurement*, *97*, 116.

Lohman, D. F. (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature.* Stanford University Calif School of Education. https://apps.dtic.mil/sti/citations/ADA075972

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In *Advances in the psychology of human intelligence, Vol. 4.* (S. 181–248). Lawrence Erlbaum Associates, Inc.

Mittring, G., & Rost, D. H. (2008). Die verflixten Distraktoren: Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests (für besser Begabte und Hochbegabte). *Diagnostica, 54*(4), 193–201. https://doi.org/10.1026/0012-1924.54.4.193

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. UTB.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Hossiep, R., Turck, D., Hasella, M. (1999). *BOMAT - advanced: Bochumer Matrizentest ; Handanweisung*.

Schmidtke, A., & Raven, J. C. (1978). *CPM Raven-Matrizen-Test: Coloured progressive matrices: Manual.* Beltz Test.

Thissen, A., Koch, M., Becker, N., & Spinath, F. M. (2018). Construct Your Own Response: The Cube Construction Task as a Novel Format for the Assessment of Spatial Ability. *European Journal of Psychological Assessment*, *34*(5), 304–311. https://doi.org/10.1027/1015-5759/a000342

Thissen, A., Spinath, F. M., & Becker, N. (2019). Manipulate Me: The Cube Construction Task Allows for a Better Manipulation of Item Difficulties Than Current Cube Rotation Tasks. *European Journal of Psychological Assessment*, 1–9. https://doi.org/10.1027/1015-5759/a000534

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

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing Reliability and Stability in Panel Models. *Sociological Methodology*, *8*, 84. https://doi.org/10.2307/270754

White, A. P., & Zammarelli, J. E. (1981). Convergence Principles: Information in the Answer Sets of Some Multiple-Choice Intelligence Tests. *Applied Psychological Measurement*, *5*(1), 21–27. https://doi.org/10.1177/014662168100500103

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence*, *30*(6), 537–554. https://doi.org/10.1016/S0160-2896(02)00086-7