Psychological Test and Assessment Modeling, 2023.65:321-338
DOI: https://doi.org/10.2440/001-0002

Conceptualization of a new "Two-way Figural Reasoning-Test"

Larissa Bartok & Klaus D. Kubinger

University of Vienna

Abstract: A concept for a so-called matrices test (a missing figure has to be deduced by reasoning, according to row- and/or column-wise logically proceeding figures in a matrix) is introduced which is new in the following ways: 1) Instead of using a typical 3 × 3 matrix design, a 5 × 5 matrix design is used, 2) in contrast to common matrices, highly redundant cells are emptied, and 3) the questioned cell is located rather at random instead of in the last row and the last column. Due to the well-known problems of a multiple-choice response format (i.e. recognizing the solution instead of producing; the phenomenon of lucky guessing; construction of plausible distractors), a free response format is used: The testee has to draw the solution by hand, whereby the necessary drawing ability is extremely low. A first draft of the *Two-way Figural Reasoning-Test* essentially measures in accordance with the Rasch model. However, testing whether the given item generating rules sufficiently explain the resulting item difficulty parameters by means of the LLTM (linear logistic test model) disclosed that some up to now non-aware (cognitive) operation components interfere with the items' difficulty. Above all, further research is needed in order to remove any not reasoning-based radicals of item difficulty – some suggestions of such radicals are given.

Keywords: matrices test, multiple-choice response format, Rasch model, item generating rules, LLTM

Author Note

Mag.[a] Larissa Bartok, BSc, University of Vienna, Center for Teaching and Learning. larissa.bartok@univie.ac.at

Prof. Klaus D. Kubinger, PhD, MSc. c/o University of Vienna, Faculty of Psychology. klaus.kubinger@univie.ac.at

Introduction

Obviously, with reference to *Thurstone*'s well-established factor reasoning (cf. Thurstone, 1938) on one hand and to *Cattell*'s generally acknowledged efforts for so-called culture-fair intelligence testing (Cattell, 1963) on the other hand, there has been a long tradition of publishing so-called matrices tests, primarily based on Raven (1938). Usually, eight cells of a [3 × 3] matrix are filled with row- and/or column-wise logically proceeding figures, the missing figure in the ninth cell (i.e. in the third row and third column) has to be determined by the testee. According to Kubinger (2023), such tests meet one of six categories of reasoning tests when crossing both *fluid* vs. *crystallized* facets (*sensu Cattell*) and *lexical* vs. *numerical* vs. *figural* contents (basically following Jäger, 1984). All of them may be assumed to measure the "ability to realize regularities and logically compelling connections in order to put them to appropriate use" (Kubinger, 2019, p. 244; translation by the authors).

From a psychometric point of view, two issues are of great importance: First, even *Raven's* world-wide published *Standard Progressive Matrices* (SPM; e.g. Horn, 2009) has proven not to meet respective standards (cf. Kubinger, Formann, & Farkas, 1991): the Rasch model does not hold, however, it's validness is essential when (only) the number of solved items are scored (see Fischer, 1995, for mathematical proof). Second, given the items of a matrices test meet the Rasch model and they were even constructed by using some item generating rules (i.e. the logic underlying the successive figures is planned in advance), a confirmation in accordance with Formann (1973) is commonly missing: do the difficulty parameters of these rules actually sufficiently explain the resulting item parameters? The latter being tested best by the so-called LLTM (linear logistic test model; Fischer, 1973; see also Fischer, 2005, as well as Kubinger, 2008, 2009). Obviously, if such a confirmation is missing, it is rather ambiguous which specific ability the test actually measures.

Moreover, it is of substantive importance that matrices tests regularly use a multiple-choice response format. Which rationalization is ever given for such a use, there is a non-zero probability for lucky guessing (see for the effectivity of certain psycho-technological response options in order to minimize that probability: Kubinger & Gottschall, 2007; Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2010; Kubinger & Wolfsbauer, 2010; Kubinger, 2014). Apart from that, the construction of distractors poses a great challenge insofar as every distractor should have, for testees with a low ability, the same plausibility of being correct (and the same plausibility as the solution itself; cf. Undeutsch, 2012). Finally, tests with multiple-choice response formats most likely only refer to the ability of recognizing the solution but do not guarantee to measure the ability of producing it.

Above all there is, indeed, the [3 × 3] *Viennese Matrices-Test* (WMT-2; Formann, Waldherr, & Piswanger, 2011), which not only fulfills the psychometric requirement that the Rasch model holds but the items also fit the LLTM (cf. Formann, 1973). However, this test uses a multiple-choice response format, though in contrast to the common format "1 out of 5", it is a "1 out of 8"-format (a single answer option of

eight given ones is correct) which at least reduces the so-called *a-priori* guessing probability (i.e. the probability to solve an item only by chance, but not with any related ability) from 1/5 = .20 to 1/8 = .125.

On the other hand, there are a few suggestions on how to avoid the multiple-choice format by applying a free response format. For instance, the computerized test [3 × 3] *Free Response Matrices* (FRM; Piskernik, 2013) asks the testee to compose the solution "pixel"-wise. Each of the eight cells of a 3 × 3 matrix is made up of 5 × 5 squares; each square is either black or white, where all black squares together make up a figure, and the white squares comprise the background. The ninth cell is fully white and the testee has to create the solution by clicking the respective squares to make them black so that the correct pattern is shown. This format is somewhat also a multiple-choice response format, although the *a-priori* guessing probability only amounts to $(½)^{25} \sim 0$. Then there is the computerized approach of DESIGMA (*Design a Matrix*; Becker & Spinath, 2014, see also Krieger, Becker, Greiff, & Spinath, 2022), where the testee needs to compose the solution using elements of a construction kit. This is in fact completely a free response format, because the number of possible combinations of elements tends to infinity. While the Rasch model holds for the items of the test FRM, this psychometric claim was not (yet) tested for DESIGMA; and both did not apply LLTM-analysis in order to check whether the conceptualized item generating rules satisfactorily explain the item difficulties.

In this paper, a test concept is presented, which is in some ways new, though obvious. It uses a free response format, has essentially stood the test of measuring reasoning in accordance with the Rasch model, and also an LLTM-analysis was applied, which by now however reveals that some non-aware (cognitive) operation components interfere with the items' difficulty. Furthermore, this test concept promises the creation of more difficult items than the commonly used 3 × 3 matrix design does.

Method

The test to be presented here is the *Two-way Figural Reasoning-Test*.

Its conceptualization is characterized by

1) being a paper-pencil test, which allows a free response format: The testee has to draw the solution by hand, whereby the necessary drawing ability is extremely low;
2) instead of using a 3 × 3 matrix design as in common matrices tests, a 5 × 5 matrix design is used;
3) also in contrast to common matrices tests, not every cell but the questioned one is filled, instead only those that deliver relevant information for finding the solution: that is, highly redundant cells are empty;

4) and finally, the element in question is located randomly in the matrix rather than strictly in the last row/last column.
5) Although not yet realized, it is also possible to query more than one cell, just because only some cells are filled.

Figure 1 gives an example of the items. This is a very easy item. According to the first row, there is the regularity "varying" (small circle – large circle – small circle – … circle – small circle) and according to the third column the regularity "proceeding" (1 circle – 2 circles – … circles – 4 circles – 5 circles). The testee must work out the figure of the cell in the third row/fourth column. If, as in common matrices tests, the third row and the fourth column were fully (redundantly) filled, the testee would not need to detect the two rules, but rather solve the item using simpler cognitive efforts: on the one hand, there are always three circles in the third row and on the other hand, there are always large circles in the fourth column.

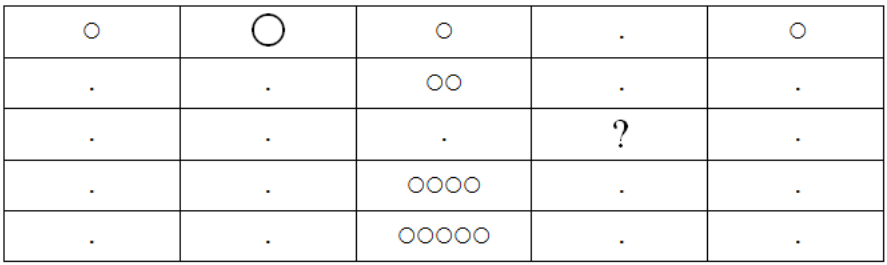

Figure 1

Instruction item of the Two-way Figural Reasoning-Test. The solution is ◯◯◯.

The item generating rules were settled with regard to Formann (1973; see also Hornke & Rettig, 1989, and Undeutsch, 2012), that is there are the following so-called radicals:

1) Number of relevant components of an item matrix's figures
2) Material properties, i.e. form, pattern, color, size, position, number, spatial arrangement
3) Type(s) of regularity (logical connection), i.e. "varying" and/or "proceeding" and/or "compounding"
4) Direction of regularities, i.e. "horizontally" and/or "vertically"
   Moreover, there is the radical
5) position of the questioned cell.

The kinds of regularity "varying", "proceeding", and "compounding" concern the following material properties:

"varying": form, pattern, color, size, position, number, spatial arrangement

"proceeding": size (continuously vs. alternately), position (a component "moves" forwards vs. backwards), number of elements of a specific component, color (continuously vs. alternately), spatial arrangement (direction of a component continuously vs. alternately altering), form (alternately altering yes vs. no)

"compounding": pattern, position, number, spatial arrangement

We give an example as an illustration – take into account that the (introduction) item given in Figure 1 has ad 1) only a single component (i.e. circle) in its 5×5 cells (figures), the number of elements of which is 1 or 2 (or 3) or 4 or 5; ad 3) the regularity "varying" of that item concerns the components' size (small vs. large), the regularity "proceeding" concerns the number of elements (always 1 element added); ad 4) in this item it happens "varying" horizontally, "proceeding" vertically.

In item 15 (Figure 2), there are ad 1) 7 components (i.e. rectangle, square, triangle, arrow [left or right directed], plus sign, white circle, and black circle), each always presented once at most (exactly three of them, a triple, in each cell). Ad 3) the regularity "proceeding" of the item refers horizontally to the component arrow's position within the triple (first, second, third). The regularity "varying" refers horizontally as well as vertically to the arbitrarily position of the five components square, triangle, plus sign, white, and black circle – thereby these variation always happens column-wise either at the second or at the third position (furthermore, the rectangle and the arrow are column-wise always in the same position, the rectangle either at the first or at the second one). Finally, the regularity "proceeding" refers also vertically to the component arrow's spatial arrangement (left or right directed), i.e. alternately. As already indicated ad 4) there are two regularities "proceeding", one horizontally and one vertically, and a regularity "varying" horizontally as well as vertically.

Based on this conceptualization, a first draft of the *Two-way Figural Reasoning-Test* has been designed (two parallel forms with a different questioned cell and consequently not the same location of filled cells; 20 items each). 192 students aged between 17 to 18 years were tested at high-schools in two countries of Austria.

The aim of the study was first to uncover whether the Rasch model holds for the given item pool (and if not, whether *a-posteriori* model validness can be established after deleting very few items); and second, if so, whether (according to the LLTM analysis) the item generating rules' difficulty parameters sufficiently explain the resulting item parameters.

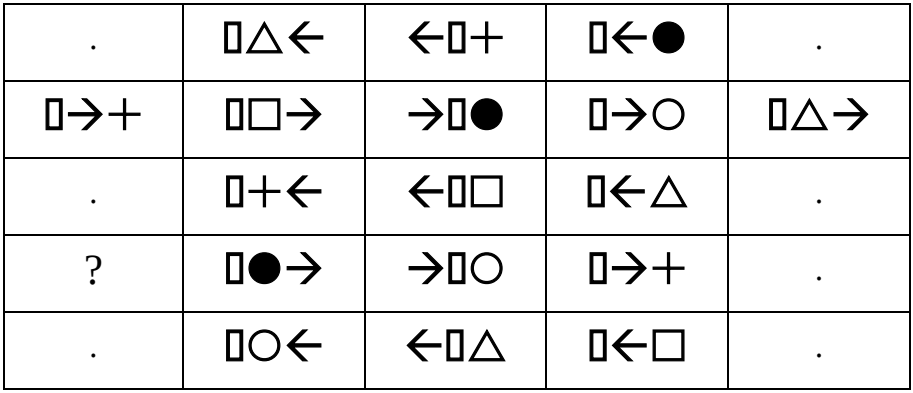

Figure 2

Item 15 of the Two-way Figural Reasoning-Test. The solution is ▯→△ *(both the first two elements of the asked triple are determined by the second row/first column, because the rectangle and the arrow are column-wise always at the same position – thereby the arrow is in the second and fourth row right directed; for the third element the square or the triangle come into question because they are missing in the fourth row, while according to the second row each component out of the five components square, triangle, plus sign, white, and black circle does row-wise occur, but the square is ruled out because according to the second, the third, and the fourth column each of the mentioned five components also occurs culumn-wise however the triangle already does in the fifth column – hence the square of the fourth row has to be there, in the fifth column).*

Rasch model calibration of the items was done in accordance with state of the art (cf. Kubinger, 2005), that is Andersen's Likelihood-ratio test (LRT) is used with several partition criteria of the calibration sample (in our case 1. score: "high-" *vs.* "low-scorers", meaning the partition in testees with a high number of solved items *vs.* testees with a low number of solved items; 2. sex: male *vs.* female testees; 3. school: students from a regular high school *vs.* students from a technical high school; 4. first language: students with German as the first language *vs.* students with a first language other than German; 5. parallel form: students tested with parallel form A *vs.* students tested with parallel form B). Given any significant LRT (we used here a comparison-wise type-I-risk $\alpha = .01$ – running five comparisons meets a study-wise type-I-risk of approximately $\alpha = .05$), items might be deleted step by step when repeating this model test until it either results in non-significance for each partition criterion, or calibration needs to be abandoned due to too many deleted items: then non-conformity of the items with the Rasch model must be found. Which item to delete can be decided with the help of Rasch's graphical model check; that check illustrates the coincidences of item parameter estimations when based on different subsamples.

The sample size has been determined according to Kubinger, Rasch, and Yanagida (2009; see also 2011). Based on their approach of testing the Rasch model by applying a three-way analysis of variance of the design $(A \succ \boldsymbol{B}) \times C$ (there is a fixed group factor *A*, a random factor *B* of testees nested within *A*, and a fixed factor *C* of items which is cross-classified with $(A \succ B)$), their effect size simulation study disclosed the following: Given the precision requirements a) a nominal significance level of $\alpha = .05$, b) an aimed-for power of $1 - \beta = .80$, and c) a defined relevant effect of item parameter difference (differential item functioning) with respect to two interesting groups of at least two times 1.00 (-0.5 instead of 0.5 in a first item and 0.5 instead of -0.5 in a second item), a sample size of $n = 101$ for each group is needed. The realized full sample size of 192 testees is negligibly smaller.

For analyzing the data with respect to the Rasch model and the LLTM, the R-package eRm (Mair, Hatzinger, Maier & Rusch, 2015) was used.

Results

Table 1 summarizes the results of Andersen's Likelihood-ratio test (LRT) with respect to the five partition criteria mentioned above.

Table 1
The Rasch model tests for 20 items of the first draft of the Two-way Figural Reasoning-Test. For the applied five criteria of partitioning the calibration sample, the results of the asymptotically χ^2-distributed Andersen's Likelihood-ratio test statistic (LRT) are given as well as the degrees of freedom (df) and the respective p-value. The results are based on 192 testees.

partition criterion	χ^2	*df*	*p*
score	22.14	19	.277
sex	24.66	19	.172
school	26.85	19	.108
first language	34.50	19	.016
parallel form	68.76	19	.000

The results show only a single significant LRT. And this concerns the partition criterion parallel form. Fundamentally, "high-" *vs.* "low-scorers" is the most powerful partition into two subsamples, given the Rasch model does not hold (cf. Kubinger, 1989); hence, our result rather indicates a problem of parallel form construction than a

problem with the test conceptualization. The graphical model check in Figure 3 reveals in particular a misfit of item 3. There, the item parameter estimations within the respective two subsamples (parallel form A *vs.* B) are opposed in a Cartesian coordinate system; ideally, the resulting dots lie on a 45° line which meets the origin – this is because, given Rasch model's validness, each item achieves the same parameter (estimation), regardless of which subsample is used. Considering the confidence ellipse, which results when the standard error of estimation is taken into account (cf. Koller, Alexandrowicz & Hatzinger, 2012) item 3 does not at all meet this 45° line. Apparently, the item parameter of item 3 is much larger for parallel form B (ordinate) than for parallel form A (abscissa), which means it is much easier in relation to the other items within parallel form B than within form A – although the parameters are called difficulty parameter, they are scaled the other way round in eRm.

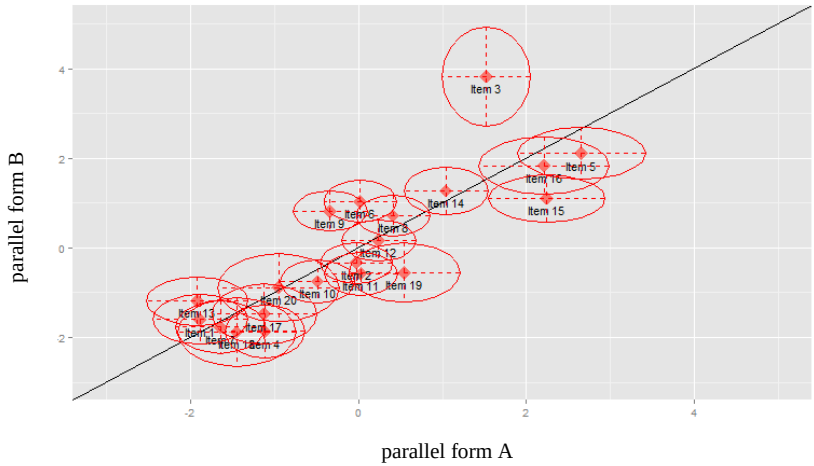

Figure 3
Graphical model check of 20 items of the Two-way Figural Reasoning-Test – item parameter estimations according to the Rasch model as opposed for students tested with parallel form A and students tested with parallel form B. For the items, not only the (estimated) item parameters are plotted against each other but the confidence ellipses are also shown. These result when the standard error of item parameter estimation is taken into account (α = .01).

Figure 4a and 4b presents item 3 in both forms, A and B. In both forms, the same components and the same regularities are of relevance. The difference is that in form B the questioned cell is in the fourth row instead of the second row, where it is in form A. With that given evidence, there are two likely explanations for the easier finding of the solution in form B. Perhaps the circumstance that the answer should already be given in the second row, as it is the case in form A, makes solving the item more

difficult, because if the testee searches for the underlying rules row by row downwards, he/she might feel pressured to produce a solution sooner, despite only having very little information. Another explanation could be that only for the given item 3, the information that in the fourth column the last component is not always a triangle is missing in form A (other than in form B). The latter shows a fault in the construction of a parallel form item. The former, on the other hand, indicates an additional radical that has to be taken into account when the difficulty of an item is designed.

.	.	.	○●●●▽	.
?	.	.	.	.
●●○●↗	●●○●●	●●○●□	●●○●▽	●●○●+
.	.	.	●○●●▽	.
.	.	.	●●●○▽	.

Figure 4a
Item 3 of the Two-way Figural Reasoning-Test in parallel form A. The solution is ●●●↗○.

Actually, deleting item 3 from the pool and re-analyzing the remaining item pool with regard to the validness of the Rasch model led again to a significant LRT (the *p*-value for each partition criterion but that for parallel form is greater than .01, for the criterion parallel form however $p = .000$). The same is true after also item 9 has been deleted ($p = .004$ for the criterion parallel form), but additionally deleting item 6 the LRT resulted in non-significance with respect to every partition criterion (see Table 2) – in Figure 3 it can be seen that the confidence ellipse of the items 9 and 6 do not meet the 45° line. Again, both items are easier in parallel form B than they are (in relation to the other items) in form A.

.	.	.	○●●●▽	.
.	.	.	●●●▽○	.
●●○●↗	●●○●●	●●○●□	●●○●▽	●●○●+
?	.	.	.	.
.	.	.	●●●○▽	.

Figure 4b
Item 3 of the *Two-way Figural Reasoning-Test* in parallel form B. The solution is ●○●●↗.

Table 2
The Rasch model tests for 17 items of the first draft of the Two-way Figural Reasoning-Test. For the applied five criteria of partitioning the calibration sample, the results of the asymptotically χ^2-distributed Andersen's Likelihood-ratio test statistic (LRT) are given, as well as the degrees of freedom (df) and the respective p-value. The results are based on 192 testees.

partition criterion	χ^2	*df*	*p*
score	12.75	16	.691
sex	15.72	16	.473
school	19.98	16	.221
first language	31.24	16	.013
parallel form	22.49	16	.128

Figure 5a and 5b shows item 9 in both forms, A and B. Figure 6a and 6b shows item 6 in both forms. Again, item 9 is easier in form B than in form A perhaps due to the fact that the questioned cell is in a row further down. However, there is also the matter of missing information in form A (as opposed to form B), namely that it becomes clear in the fifth column that the first component is not always a black circle. Both explanations for why finding the solution in form B is easier do not apply to item 6. Perhaps it is much easier to draw a single rectangle (form B) than five triangles (form A), which again would mean an additional radical needs to be taken into account, when the difficulty of an item is designed.

●◴△▯□	●◴△□▯	●◴□△▯	●□◴△▯	.
.	.	●◴⊠△▯	.	.
.	.	●◴⊡△▯	.	?
.	.	●◴▣△▯	.	.
.	.	●◴☑△▯	.	.

Figure 5a:
Item 9 of the Two-way Figural Reasoning-Test in parallel form A. The solution is ⊡●◴△▯.

.	●◴△□▯	●◴□△▯	●□◴△▯	□●◴△▯
.	.	●◴⊠△▯	.	.
.	.	●◴⊡△▯	.	.
?	.	●◴▣△▯	.	.
.	.	●◴☑△▯	.	.

Figure 5b
Item 9 of the Two-way Figural Reasoning-Test in parallel form B. The solution is ●◴△▯▣.

▲	△△	▯▯▯	◇◇◇◇	●●●●●
.	●●	▲▲▲	△△△△	.
.	◇◇	●●●	▲▲▲▲	.
.	▲▲	◇◇◇	▯▯▯▯	?
.	▯▯	△△△	●●●●	.

Figure 6a

Item 6 of the Two-way Figural Reasoning-Test in parallel form A. The solution is △△△△△.

▲	△△	▯▯▯	◇◇◇◇	●●●●●
?	●●	▲▲▲	△△△△	.
.	◇◇	●●●	▲▲▲▲	.
.	▲▲	◇◇◇	▯▯▯▯	.
.	▯▯	△△△	●●●●	.

Figure 6b
Item 6 of the Two-way Figural Reasoning-Test in parallel form B. The solution is ▯.

At the end, the graphical model check results as shown in Figure 7. The item parameters for the remaining 17 items lie between -1.54 and 2.50, which means a rather small range from experience.

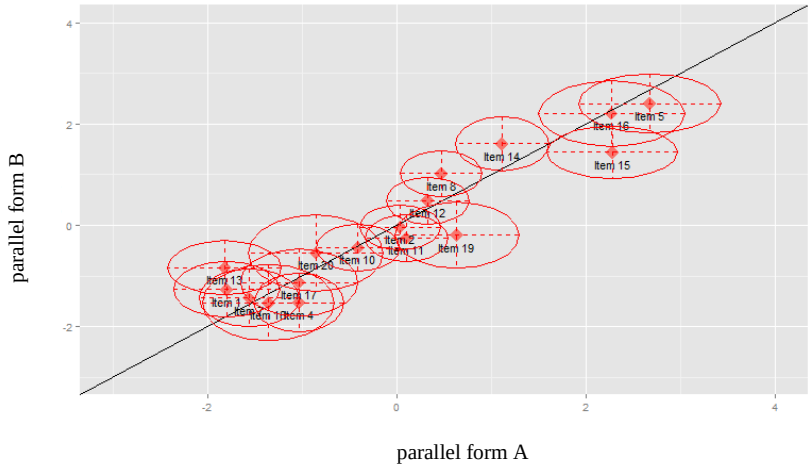

Figure 7
Graphical model check of the remaining 17 items of the Two-way Figural Reasoning-Test – item parameter estimations according to the Rasch model as opposed for students tested with parallel form A and students tested with parallel form B. For the items not only the (estimated) item parameters are plotted against each other but the confidence ellipses are also shown. These result when the standard error of item parameter estimation is taken into account (α = .01).

There has been a second study in order to test the Rasch model's validness for the *Two-way Figural Reasoning-Test*[1]. Analyses fundamentally confirmed the observation that the question mark in different cells may unintentionally offer different information for the same item, which either makes it easier or harder to realize the regularities and logically compelling connections of the figures' components.

Analysis of our data by the LLTM were based on the described item generating rules and radicals, respectively. This model decomposes the Rasch model's item parameter σ_i, $i = 1, 2, \ldots k$ by using a linear combination of some elementary operation parameters η_j ($j = 1, 2, \ldots p < k$); that is $\sigma_i = \sum_j^p q_{ij}\eta_j$, where q_{ij} weights these operation parameters according to a given hypothesis. The so-called structure matrix ((q_{ij})) is shown in Table 3. Bear in mind that some radicals (operation parameters) had to be deleted in advance in order to get a full rank of this matrix, which is necessary for the estimation of these parameters.

[1] This study was carried out for the Master Thesis of *Benedikt Winter* and supervised by the second author as the responsible university advisor. For details see Winter (2016).

Table 3

The structure matrix ((q_{ij})) of the item difficulties' linear combination of elementary operation parameters of the Two-way Figural Reasoning-Test due to the described item generating rules (empty cells mean $q_{ij} = 0$).

item	operation: proceeding size alternately	proceeding number of elements 1 or 2	proceeding number of elements >2	proceeding direction alternately altering	proceeding direction continuously altering	proceeding position forwards	proceeding form alternately altering	varying form	varying position	varying number	varying pattern	varying spatial arrangement	compounding any
1	1	1											
2		1		1	1								
4			1	1									
5						1	1						1
7	1				1								
8		1										1	
10		1		1			1						1
11		1						1					
12										1			1
13		1						1					1
14					2						1		
15					1			2					
16						2							
17			1				1						
18			1				1						
19	1								1				
20	1											1	

The hypothesis that the item generating rules' difficulty parameters explain sufficiently the resulting 17 (Rasch model) item parameters has been tested by the pertinent goodness-of-fit likelihood ratio test (cf. e.g. Kubinger, 2008). The asymptotically χ^2-distributed test statistic resulted as $\chi^2 = 49.50$ ($df = 3$), which leads to $p = .000$. This means there are some non-aware cognitive operation components interfering with the items' difficulty.

Discussion

First of all, the *Two-way Figural Reasoning-Test* stood the test. The Rasch model holds for the constructed item pool – given no parallel form is used. Most probably, similar items that were constructed along the same item generating rules would also meet the Rasch model's validness.

However, the LLTM analysis disclosed clearly that the hypothesized operation parameters with regard to the item generating rules do not (at all) suffice to explain the items' difficulty (parameters). A reason might be that because of very few used items (i.e. 17), the estimation of each designed operation parameter is only based on very few items. As a consequence of this, arbitrary item components may essentially bias the parameter estimation. That is, follow-up studies have to use a substantially greater number of items (60 to 100) to get conclusive parameter estimations. It might also be worthwhile to hypothesize that some item components (generating rules and radicals, respectively) do not affect additively but rather "multiplicatively", meaning some compounded effects should be represented by a separate operation parameter.

However, it is more likely that content reasons are fundamentally responsible for the result of the LLTM analysis. Alone the given parallel form item effects proved that there are radicals with hardly any concern to reasoning. While the use of parallel forms could either be abandoned or respective items are to simply be constructed by changing symbols, such radicals will invalidate the aimed-for measurement of reasoning – though their difficulty might be appropriately taken into account. Ultimately, the position of the questioned cell must not influence the problem-solving behavior. If follow-up studies confirm our suggestion, that the solution becomes easier in the case the questioned cell is in the fourth row instead of in the second one (and the like), then the questioned cell should simply not be located in the first, second and maybe third row. Otherwise, some personality trait (e.g. impulsivity) runs the risk of contaminating the test-score. Furthermore, if follow-up studies support our assumption that solutions are generally harder to find if not only a single symbol is to be drawn but rather multiple symbols, then solutions should always require as little drawing effort as possible – or the free response format is, after all, not suitable. Coming to the point, specific research is needed for each item in order to check for the occurrence of any not reasoning-based radical of the item difficulty. We suggest applying the method of thinking aloud for further studies. In this exploratory method, the subject is asked to

work on the items while he/she verbally expresses all thought processes and action strategies. Maybe any non-aware cognitive operation components interfering with the given item generating rules could be detected by these means; and it is most likely that effects due to certain energetic-motivational factors (e.g. the efforts for drawing) would be recognized in this manner – apart from the facility to detect design flaws of an item, especially the presence of an alternative solution according to unaware, underlying rules.

Despite the open questions raised, the conceptualization of the *Two-way Figural Reasoning-Test* seems encouraging. And this test concept probably counteracts the experience (Preckel, 2003) that the commonly used 3×3 matrix design rather limits the difficulty of items.

References

Becker, N. & Spinath, F. M. (2014). *DESIGMA – Advanced (Design a Matrix)*. Göttingen: Hogrefe.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1995). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models* (pp. 15–38). New York: Springer.

Fischer, G. H. (2005). Linear logistic test models. *Encyclopedia of Social Measurement, 2*, 505-514.

Formann, A. (1973). *Die Konstruktion eines neuen Matrizentests und die Untersuchung des Lösungs-verhaltens mit Hilfe des linearen logistischen Testmodells* [The construction of a new matrices-test and the analysis of the testees‘ solving behaviour by the linear logistic test-model]. Unveröffentlichte Dissertation [unpublished doctoral thesis], Universität Wien [University of Vienna].

Formann, A. K., Waldherr, K., & Piswanger, K. (2011). *Wiener Matrizen-Test 2, WMT-2* [Viennese Matrices-test, WMT-2]. Göttingen: Hogrefe.

Horn, R. (ed.) (2009). *Standard Progressive Matrices (SPM)* (2nd ed.). Frankfurt/M.: Pearson.

Hornke, L. F., & Rettig, K. (1989). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen [Item generating rules with regard to basics of cognition psychology]. In K. D. Kubinger (ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions] (pp. 140-83). Munich: PVU.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Research on the intelligence structure: competing models, new progresses and outlooks]. *Psychologische Rundschau, 35*, 21-35.

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). Das Rasch-Modell in der Praxis – Eine Einführung in eRm [The Rasch model in practice – an introduction in eRm]. Vienna: facultas (UTB).

Krieger, F., Becker, N., Greiff, S., & Spinath, F. M. (2022). *DESIGMA – Standard (Design a Matrix)*. Göttingen: Hogrefe.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions] (pp. 19-83). Munich: PVU.

Kubinger, K. D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*, 377-394.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50*, 311-327.

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement, 69*, 232-244.

Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester" Multiple-Choice-Prüfungsaufgaben [Expert report: How to construct multiple-choice examination tasks to stand up in court]. *Psychologische Rundschau, 65*, 169-178.

Kubinger, K. D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (eds.), *Principles and methods of test construction* (pp. 104-119). Göttingen: Hogrefe.

Kubinger, K.D. (2019). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens* (3rs ed.) [*Psychological Assessment – Theory and Prcatice of Psychological Comsulting*]. Göttingen: Hogrefe.

Kubinger, K. D. (2023). Guest Editorial: Promising reasoning test ideas not yet published. Special issue: *Promising reasoning test ideas not yet published. Psychological Test and Assessment Modeling, 65*, 315-320.

Kubinger, K. D., Formann, A.K. & Farkas, M.G. (1991). Psychometric shortcomings of Raven´s Standard Progressive Matrices (SPM) in particular for computerized testing. *European Review of Applied Psychology, 41*, 295-300.

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science, 49*, 361-374.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*, 111-115.

Kubinger, K. D., Rasch, D. & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly, 51*, 370-384.

Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17,* 321-333.

Kubinger, K. D. & Wolfsbauer, C. (2010). On the risk of certain psycho-technological response options in multiple-choice tests: does a particular personality handicap examinees? *European Journal of Psychological Assessment, 26*, 302-308.

Mair, P., Hatzinger, R., & Maier M. J. (2015). eRm: Extended Rasch Modeling. 0.15-5. http://erm.r-forge.r-project.org/

Piskernik, B. (2013). *FRM: Free Response Matrices.* [Software and Manual: Version 51 – Revision 5, 2018]. Mödling: Schuhfried.

Preckel, F. (2003). Diagnostik intellektueller Hochbegabung [Assessment of High Ability]. Göttingen: Hogrefe.

Raven, J. C. (1938). Standardization of progressive matrices. *British Journal of Medical Psychology, XIX*, 137-150.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Undeutsch, N. (2012). Schlussfolgerndes Denken figural: Der Färbige Matrizentest [Reasoning figural: The colored matrices test]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests" [Viennese authors collective]) (eds.), *Self-Assessment: Theorie und Konzepte* (pp. 143-152). Lengerich: Pabst.

Winter, B.O.J. (2016). *Dimensionalitätsvergleich der Reasoning-Tests Gleichungen, Figurale Analogien und Figurales Zweidimensionales Reasoning* [The factorial structure of the reasoning tests *Equations*, *Figural Analogies* and *Two-way Figural Reasoning-Test*]. Unpublished master thesis, University of Vienna, Austria.