Research Article

Benedict C. O. F. Fehringer*

Spatial thinking from a different view: Disentangling top-down and bottom-up processes using eye tracking

<https://doi.org/10.1515/psych-2020-0105>

Received May 29, 2020; accepted June 22, 2020

Abstract: The goal of the present study was to investigate the potential of gaze fixation patterns to reflect cognitive processing steps during test performance. Gaze movements, however, can reflect top-down and bottom-up processes. Top-down processes are the cognitive processing steps that are necessary to solve a certain test item. In contrast, bottom-up processes may be provoked by varying visual features that are not related to the item solution. To disentangle top-down and bottom-up processes in the context of spatial thinking, a new test (R-Cube-Vis Test) was developed and validated explicitly for the usage of eye tracking in three studies as long and short version. The R-Cube-Vis Test measures visualization and is conform to the linear logistic test model with six difficulty levels. All items of one level demand the same transformation steps to solve an item. The R-Cube-Vis Test was then utilized to investigate different gaze-fixation-based indicators to identify top-down and bottom-up processes. Some of the indicators were also able to predict the correctness of the answer of a single item. Gaze-related measures have a high potential to reveal cognitive processing steps during solving an item of a given difficulty level, if top-down and bottom-up processes can be segregated.

Keywords: spatial thinking, nonverbal intelligence, visualization, psychological testing, eye tracking, cognitive processing steps, top-down and bottom-up processing, Rubik's cubes

Introduction

Psychological diagnostics cover the testing of someone's abilities or personality characteristics that can lead to informed decisions in the present or about future events (e.g., Swets, Dawes, & Monahan, 2000). Especially in performance testing, it might be interesting to get insights into the cognitive steps that lead to the given answer of a certain task. On the one hand, such information could be used to detect inappropriate solving strategies, which might be important if the test performance is poor. If the knowledge about appropriate solving strategies is not intended to be measured by the test, the information about the usage of inappropriate strategies could be used to, e.g., reevaluate the test or to teach appropriate solving strategies for a follow-up test. On the other hand, the knowledge about such cognitive steps might be able to differentiate between participants with high and low ability even if the items have ceiling or floor effects and/or have a high guessing probability.

However, most of the common performance tests are based on a single measure per item, accuracy (or a derivate) or reaction times. Information about *how* participants solve certain items is usually not

---

*Corresponding author: Benedict C. O. F. Fehringer, University of Mannheim, Department of Psychology of Education A5 6, 68159 Mannheim, Germany, E-mail: b.fehringer@uni-mannheim.de

considered. Only if the tasks are more complex such as in the MicroDyn test (Greiff, Wüstenberg, & Funke, 2012), in which participants have to solve linear structural equation models for simple sample scenarios, sequences of single mouse clicks can be recorded to identify solving strategies (e.g., Meißner, Greiff, Frischkorn, & Steinmayr, 2016; Schweizer, Wüstenberg, & Greiff, 2013). If the tasks are short and can only be answered as, e.g., correct or incorrect, only accuracy and reaction times are usually accessible. Current research addresses this issue by taking reaction times into account as a moderating factor if accuracy is the main measure. For example, van der Linden (2007) proposed a hierarchical model that includes accuracy and reaction times per item on an individual and population level. This allows evaluating the accuracy of an answer by the time needed to produce this answer. However, a single reaction time value can only be an approximate indicator for the cognitive processes performed during the solving of the item. Generally, it cannot be answered whether longer or shorter reaction times indicate a reasonable solving behavior. An alternative approach, the multinomial processing tree model (Erdfelder & Buchner, 1998), attempts to decompose cognitive processes based on accuracy but still utilizes a single value per item. Besides these traditional measures, eye tracking provides alternative measures to access cognitive processes due to the many data points that can be gathered per item.

The goal of the presented studies was to test whether such eye tracking information has the potential to provide information additional to accuracy and reaction times in order to describe someone's ability and to evaluate the solving behavior. This research question was analyzed in the well-established domain of spatial thinking. The standard spatial thinking tests, however, are limited if they are used in eye tracking research due to some restriction, such as heterogeneous stimulus materials and overlapping of relevant item features (see below). To this end, a new test was developed and validated for the explicit usage of eye tracking.

The remainder of the paper is organized as follows: Section 2 introduces the construct of spatial thinking and provides an overview over current eye tracking research in this domain. The drawbacks of the current standard tests are also considered. In Section 3, the newly developed spatial thinking test, the R-Cube-Vis Test, is described in detail, including first evidence for validation aspects. Study 1 to 3 (Sections 4 – 6) analyze further aspects of validity of the R-Cube-Vis Test as a long and a short version. The fourth study (Section 7) demonstrates the usage of the R-Cube-Vis Test in eye tracking research by disentangling bottom-up and top-down processes and by applying simple and sophisticated fixation-based measures. The paper ends with a general discussion.

Spatial thinking

Spatial thinking plays an important role as a nonverbal intelligence component (e.g., Paivio, 2014) in developing specialized knowledge in STEM (science, technology, engineering, and mathematics) domains, in addition to mathematical and verbal abilities (Wai, Lubinski, & Benbow, 2009). Furthermore, it influences learning from visualizations (Höffler, 2010; Münzer, 2012, 2015) and is related to vocational interests (Pässler, Beinicke, & Hell, 2015) as well as to imagery preferences and creativity (Kozhevnikov, Kozhevnikov, Yu, & Blazhenkova, 2013). Moreover, spatial thinking ability is a critical factor in choosing solving strategies in spatial tasks (Wang & Carr, 2014) and influences how people benefit from supporting aids in learning from visualizations (Münzer, 2012).

The construct of spatial thinking is considered to be heterogeneous. According to the factor-analytic approach, it consists of different factors that can be measured with specific tests (see Hegarty & Waller, 2005). The number of the spatial thinking factors differs from two to five depending on how broad the construct or single aspects are defined (e.g., Carroll, 1993; Lohman, 1988; McGee, 1979; Michael, Guilford, Fruchter, & Zimmerman, 1957). One of the most extensive works came from Carroll (1993), who reanalyzed more than 90 studies. His five-factor structure could also be confirmed by Burton and Fogarty (2003) and forms the basis of the present studies. The first factor is visualization referring to the “the ability to manipulate or transform the image of spatial patterns into other arrangements” (Ekstrom, French, Harman, & Dermen, 1976b, p. 173). The second factor is spatial relations and means “the ability to perceive spatial patterns or to maintain

orientation with respect to objects in space” (Ekstrom et al., 1976b, p. 149). In contrast to visualization, where the configuration of object parts is changed, spatial relations refers to the manipulation of the whole object (Ekstrom et al., 1976b). These two factors can be interpreted as the ends of a continuum (Pellegrino, Alderton, & Shute, 1984) with complex and power tests on the one end (visualization) and simple and speed tests on the other end (spatial relations). The further three spatial thinking factors of Carroll (1993) cover the spatial abilities to recognize objects under different conditions (Closure Speed, Flexibility of Closure, Perceptual Speed) and to memorize objects (Visual Memory).

The present work focuses on the main factor, visualization, and contrasts it with the second factor, spatial relations. Both factors are not distinct but correlated with each other, $.40 \leq r \leq .57$ (Kozhevnikov & Hegarty, 2001; see Kozhevnikov, Hegarty, & Mayer, 2002; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). Correlations between different spatial relations tests are between $.34 \leq r \leq .62$ (see Hegarty & Waller, 2004; Kozhevnikov et al., 2002; Kozhevnikov & Hegarty, 2001; Vandenberg & Kuse, 1978) and correlations between different visualization tests are between, $.51 \leq r \leq .63$ (Blajenkova, Kozhevnikov, & Motes, 2006). Both spatial thinking factors are also correlated with mathematical and verbal abilities, although some of these correlations are very small and therefore irrelevant, $.07 \leq r \leq .52$ and $.17 \leq r \leq .65$ (see, Fennema & Sherman, 1977; Geary, Saults, Liu, & Hoard, 2000; Gunderson, Ramirez, Beilock, & Levine, 2012; Hegarty & Kozhevnikov, 1999; Kaufman, 2007; LeFevre et al., 2010).

As was generally pointed out for performance tests before, standard tests for visualization provide only accuracy values and, therefore, are restricted to describing someone’s ability. Potential solving strategies that may vary between participants cannot be obtained with these tests (Hegarty & Waller, 2005). Different participants might even solve task items in different manners, depending on their abilities and preferences (see Kyllonen, Lohman, & Snow, 1984; Kyllonen, Lohman, & Woltz, 1984). These drawbacks are illustrated in the following two standard tests for spatial thinking, the Paper Folding Test (PFT, Ekstrom, French, Harman, & Dermen, 1976a, Figure 1) and the Mental Rotation Test (MRT, Vandenberg & Kuse, 1978), which were both conducted in the presented studies for validation purposes.

Figure 1: Sample item of the Paper Folding Test (PFT). The task is to decide which of the paper sheets on the right side (A – E) matches the unfolded paper sheet on the left side. Here, the correct answer is C.

According to Ekstrom et al. (1976b), a typical visualization test is the Paper Folding Test (PFT, Figure 1), where participants have to mentally unfold a perforated paper sheet. Another test that was invented as a visualization test is the Mental Rotation Test (MRT, Figure 2), which, however, does not fulfill all aspects of the definition of visualization. In the MRT, participants have to mentally rotate, but neither to transform nor to manipulate, 3D objects that were originally invented by Shepard and Metzler (1971). Although Carroll (1993) assigned the MRT as *Block Rotation Tasks* to the visualization tests, he pointed out that “block rotation tasks are likely to appear on factor SR [spatial relations] when they are simple and highly speeded.” (p. 323) Therefore, Pellegrino et al. (1984) placed the MRT in the middle of the described continuum between visualization and the second spatial thinking factor, spatial relations. For the presented studies, the redrawn version of Peters et al. (1995, MRT-A) was utilized.

Figure 2: Example task of the Mental Rotation Test in the redrawn version (MRT-A, Peters et al., 1995). Two out of the four objects on the right side have to be identified that represent a rotated version of the object on the left side. Here, the first and third object are the correct ones.

In the present context, a strategy refers to a certain sequence of different cognitive processes performed to solve a task. For example, Kyllonen, Lohman, and Snow (1984) described two solving strategies for the PFT. In the visualization strategy, participants solve tasks by mentally unfolding the left paper sheet, i.e., they cognitively process each item in the intended manner. In contrast, participants who use the analytic strategy solve these tasks by deriving logical conclusions without mentally transforming or manipulation the shown figures. The inspection of the second figure on the left side (Figure 1), for example, would lead to the conclusion that all perforations have to be on the left side of the sheet. Therefore, options B, D, and E can be excluded. Furthermore, the observation that the perforation goes through the upper and lower part of the paper, leads to the conclusion that A can also be excluded. Hence, the correct answer is C. Given these different approaches to solving these items, standard tests of visualization that are purely based on accuracy measures can fail to measure the true construct of visualization if participants actually use, e.g., analytic solving strategies. Therefore, further information about how a participant has solved a certain task is needed to derive sophisticated descriptions of someone's ability.

A promising source of information about cognitive processes taking place while participants are solving a visual task is eye tracking. Eye-tracking methods deliver gaze-related measures such as gaze fixations that are “pauses over informative regions of interest” (Salvucci & Goldberg, 2000, p. 71). The basic assumption of eye tracking research is the “eyes’ tendency to fixate the referent of the symbol that is ‘at the top of the stack’ [of the cognitive processes].” (Just & Carpenter, 1976, p. 477) This so-called eye mind hypothesis assumes that the parts of an object are cognitively processed at the same time as they are fixated by the eyes. This holds at least for cognitive tasks with rapid operations (50 – 800msec, according to Just & Carpenter, 1976).

There are only a small number of studies that use eye tracking as an additional source of information during the performance of spatial thinking tests. An important difference between these studies is the different size of areas of interest (AOIs). AOIs are defined regions on the stimulus materials in eye tracking research. Typically, distinct fixations within an AOI are not differentiated and are interpreted as equal. Only the number and duration of fixations on certain AOIs are usually reported. Fixations on specific details within these AOIs are not considered. Therefore, the definition of AOIs (with respect to its scaling) determines the resolution of the resulting gaze pattern. Fine-scaled AOIs produce more detailed information (higher resolution) than large-scaled AOIs (lower resolution).

One of the earliest studies that postulated a three-phases cognitive processing model for visualization showed characteristic gaze patterns for the three phases using large-scaled AOIs (Egan, 1979). The AOIs definition discriminated only between the whole objects or figures that had to be compared. Details of these objects were ignored. Egan (1979) was able to describe a robust viewing pattern by using this AOI definition, but the resulting model is very general and not specific. The first phase is the Search-phase, in which participants compare objects pairwise. In the second phase, the Transform-phase, participants try to transform non-matching elements of the objects until they match. In the third phase (Confirm-phase), participants compare other corresponding object elements to decide whether the objects are equal except for the transformations. The first and third phases were characterized by fixation switches between two objects, looking for matching parts. Furthermore, it appeared that the time depended on the discriminability of the two objects (Carpenter & Just, 1978). The second phase contained longer fixations that went back and forth between corresponding parts. A similar discrimination into three phases (search, transformation/

rotation, conformation) could also be found for pairs of the Shepard-Metzler-figures (Just & Carpenter, 1976). Although Just and Carpenter (1976) defined AOIs for *parts* of the presented objects by considering fixations on single arms of the Shepard-Metzler-figures (see Figure 2), the resulting gaze patterns were also only generally described due to the simple stimulus materials. In addition, the resulting fixation sequences were manually segmented and annotated with respect to the three phases, which limits objectivity.

Another application example utilizing large-scaled AOIs comes from Snow (1980), who examined inter-individual differences. To this end, he analyzed the PFT results of 48 high-school students representing extreme groups with regard to their abilities in crystallized and fluid intelligence (Cattell, 1971). In his eye tracking study, Snow (1980) considered gaze paths based on large-scaled AOIs, where each AOI represented one single figure (e.g., there were seven AOIs defined for the item presented in Figure 1). Fixations on specific features of the figures, such as perforation holes or specific areas, were not considered. However, Snow (1980) was able to find differences between both ability groups. Low-ability participants seemed to have problems in inhibitory or control mechanisms while they were checking the answer alternatives. High ability participants were able to better sustain their item analysis.

In another study, the gaze movements on fine-scaled AOIs of complex items were analyzed. Ivie and Embretson (2010) tracked gaze movements of ten participants while they solved the Revised Minnesota Paper Form Board Test (RMPFBT, Likert & Quasha, 1970). An RMPFBT item shows six figures in three rows and two columns. The target figure is placed in the upper left corner and shows randomly distributed separated parts of a simple form, e.g., a disc. All other five figures show the target form, the disc, not separated but with different partitions drawn on it. Participants have to decide which of the five figures contains the same partition as the target figure. Ivie and Embretson analyzed the data qualitatively because of the small sample size ($N = 10$) with respect to differences in participant groups (low ability, high ability, and guesser) as well as differences between the items. For their analysis, they used fine-scaled AOIs on single features of the six figures per item. Participant groups differed in the time they spent on analyzing the shown figures, or in the number fixations on features (e.g., edges) of these figures. However, specific gaze patterns in order to derive cognitive processing steps could only be found by visually analyzing the eye tracking videos. The analysis of the gaze behavior for each item over all participants revealed individual gaze patterns for each item with respect to the mean proportional time that was spent on certain item features. However, it was impossible to derive general viewing patterns for all items.

In the presented studies, an item consists of various figures (e.g., seven figures in a PFT item, Figure 1; five figures in a MRT item, Figure 2). AOIs were defined either one for each figure (large-scaled) or as fine-scaled AOIs for different features of a single figure. The results showed that fixation patterns based on large-scaled AOIs or on fine-scaled AOIs for simple items (e.g., of the MRT) are only able to describe simple cognitive processing models. Fixation patterns based on fine-scaled AOIs of complex items (e.g. of the RMPFBT or the PFT), however, would potentially be able to deliver more sophisticated cognitive processing models. But fine-scaled AOIs resulted in noisy data due to visual features of the items that influenced the gaze movement but were not related to the demanded cognitive processing steps. Together, the studies show that there exists a trade-off between the complexities of the items, the appropriate size of the AOIs that define the smallest fixation unit (item figures vs. single features of item figures) and the complexity of the derived cognitive processing model. The less complex the items, the smaller can the AOIs be defined while still arriving at meaningful results. Whereas the analysis of single features of the item figures was appropriate for the simple items of the Shepard-Metzler-figures, this was not the case for the more complex items of the RMPFBT (Ivie & Embretson, 2010). Both combinations, simple items with small AOIs (e.g., Just & Carpenter, 1976; Nazareth, 2015) and more complex items with larger AOIs (e.g., Snow, 1980) seem to be only able to define simple cognitive processing models consisting of, e.g., three phases, or to distinguish between two strategies or performance groups. For more complex items that measure the spatial thinking factor visualization more complex cognitive processing models would be necessary for an appropriate representation. However, such models could only be established if small AOIs for single features of item figures can be meaningfully interpreted.

The current tests, with their complex items, especially the PFT as a visualization test, are limited at this point due to different reasons. (1) The comparison of gaze patterns between items is hampered by the

heterogeneous item sets regarding their visual appearance. Different items differing in too many features lead to gaze patterns that are hard to interpret (e.g., Ivie & Embretson, 2010). Although the recorded gaze patterns are often interpreted as a result of purely task-driven visual attention (top-down process), it can be expected that the gaze movements reflect an interplay between top-down and bottom-up processes (such as pop-up effects, see Posner, 1980; Theeuwes, 2010). Therefore, the fixation pattern recorded for a single item is influenced by two sources: A) Top-down processes, which cover the cognitive processing steps to solve the items that result from the necessary transformation steps. Here and in the following text, “transformation steps” refer to the intended solving steps, e.g., the mental unfolding of the left sheet of an item of the PFT, which has to be distinguished from the actual cognitive processing steps. The second source of influence are B) bottom-up processes, which address gaze movements driven by visual item features that are independent of these necessary transformation steps and occur only due to irrelevant characteristics of the item’s visual appearance. In the following text, (parts of) fixation patterns resulting from top-down processes referring to necessary transformation steps are called *type-specific*. (Parts of) Fixation patterns that are driven by irrelevant visual item features (bottom-up processes) are called *item-specific*. In heterogeneous item sets, both effects cannot be disentangled and, hence, the resulting fixation patterns are hard to interpret. (2) In some cases, it is not possible to decide which item feature is cognitively processed (considering the fixation) because of the visual overlap of relevant item features. For example, in the third figure from the right shown in the MRT item (Figure 2), one edge is lying over another edge. Also, if a participant is fixating on the upper part of the second figure from the left of the PFT item shown in Figure 1, it is not possible to decide whether the participant is thinking about the folding or the perforation. (3) The presentation of many figures within a single item leads to complex gaze patterns that are hard to interpret, e.g., the PFT shows seven to nine figures per item, and the MRT shows five figures per item. (4) Some tests are presented as paper-pencil versions that limit the accuracy of the gaze estimations because of the used eye tracking hardware. In paper-pencil versions, gaze patterns had to be recorded by mobile eye trackers that would often result in less accurate gaze estimations than remote eye trackers due to different hardware components, parallaxing and slippage of the head-worn eye tracker (Holmqvist et al., 2011). Computer-based test versions would overcome these issues by using remote eye trackers.

Hence, new testing materials are needed that overcome these restrictions in order to be able to gain insights into the cognitive processes involved in solving items of spatial thinking tests for the visualization factor. To this end, the R-Cube-Vis Test was developed and validated (Study 1 to 3). However, once these materials existed, the next challenge would be the identification of fixation patterns of specific item characteristics and the decision whether these patterns indicate cognitive processes that are relevant to solve the demanded task or whether these patterns reflect idiosyncratic characteristics of the item based only on its visual appearance. The aim of Study 4 was to demonstrate which fixation-based indicators might be promising to reveal cognitive processes, how they can differentiate between indicators that are related to task demands or task irrelevant item characteristics, and whether these analyzed indicators are able to indicate the correctness of a single answer.

A new test for visualization

Regarding the four restrictions of standard visualization tests presented in Section 2, four requirements were formulated as a final solution to overcome these issues: (1) The test must contain many homogenous items of one task type (homogenous group) to differentiate between eye movements that differ between items of a single homogenous group (item-specific characteristics) and eye movements that are indicative for a specific homogenous group (type-specific characteristics). Type-specific eye movements refers to movements that correspond to the cognitive processes to solve the item, whereas item-specific eye movement refers to gaze patterns that result from the visual appearance of the item but are independent of the demanded cognitive processes to solve the item. Therefore, “homogenous” items should be equal with regard to difficulty and their construction in such a way that the specific transformations to solve items of the same homogenous group should be the same for all items. On the other side, one item should not be

solved by just remembering the answers of previous items of the same homogenous group. To this end, all items should differ in their visual appearance. Further factors that might influence the gaze movements (such as mind wandering, e.g., Uzzaman & Joordens, 2011) that result as apparent random patterns will be controlled by considering the average over items. (2) On each item figure, all relevant areas should be visually separated. If the area of a specific fixation is detected, it should be unambiguous with respect to the specific part of the item that is fixated, i.e., there should be no visually overlapping parts. (3) The items should be built by only a few figures to avoid too complex patterns of movement that are hard to interpret. (4) The test should be computer-based to guarantee standardized testing and the best possible measuring results of the applied eye tracking hardware.

New stimulus materials

Based on the formulated four requirements, a new test, the R-Cube-Vis Test, was developed in accordance with the definition of visualization that describes it as “the ability to manipulate or transform the image of spatial patterns into other arrangements” (Ekstrom et al., 1976b, p. 173, see also Section 2). Cubes as stimulus materials are common in spatial thinking tests (e.g., Cube Comparison test, Ekstrom et al., 1976a, Figure 8). However, the cubes in the existing tests are rotated as a whole, which is not in accordance with the definition of visualization that demands a manipulation or transformation of the image. The item of the R-Cube-Vis Test fulfill this criterion by movable elements of the cubes, such as in the Rubik’s cubes. The participants are asked to mentally transform single elements of one cube to get the other cube. Regarding the four requirements for eye tracking usage, the items were constructed in a way that many items of the same kind could be generated. The items are simply structured, they show all relevant parts without visual overlap, and they can be presented by a computer. Hence, the R-Cube-Vis Test is expected to measure visualization and to be utilized for eye tracking.

In the first step, a preliminary study was conducted to define the item characteristics belonging to a certain difficulty level and to select single items for the final test. All items of the R-Cube-Vis Test use Rubik’s cubes as figures (Figure 3). Each item shows two colored cubes, one target cube on the left side and one transformed/rotated cube on the right side. The task is always to decide whether elements of the right cube can be rotated such that the left cube results (possible item, Figure 3) or not (impossible item, Figure 5).

All elements of the right cube in the R-Cube-Vis Test can be potentially rotated in two directions and about three axes like in the original Rubik’s cube. Each cube consists of the same six colors (blue, brown, green, red, white, and yellow). However, there might be fewer colors visible, because only three sides are shown. The perspective of each cube shows all three sides in the same size. The final test materials consist of six difficulty levels (Figure 3). The presented cubes differ in their size regarding the number of visible pieces. If the cubes show 3 x 3 pieces on each side, the size of this cube is labeled as “size 3”. If 4 x 4 pieces are shown on each side, the cube has “size 4”. Furthermore, they differ also in the number of the rotated elements (1 vs. 2) and how two elements are rotated (parallel, p vs. crossed, c). The levels are labeled accordingly. The two most difficult levels (Level 4.2.c1 and Level 4.2.c2) differ with respect to the visible cross and the visible cross point (Figure 4).

As can be seen in Figure 5, the impossible items for each level were created to look structurally like the possible items of the same level. They always had the same size and the same number of changed elements. The impossible items were created from specific possible items of the respective level, which were not used in the test. These possible items were switched into impossible items by changing the coloring and the position of the rotated elements. However, one cannot trace back the original possible item from an impossible item, because each impossible item can be reached from different possible items. The impossible items of both most difficult levels are indistinguishable between both levels, because the possible cubes of both levels are structurally equivalent with the same size, the same number of rotated elements and the same kind of rotation.

Figure 3: Possible sample items of the six difficulty levels of the R-Cube-Vis Test.

Figure 4: Sample of possible items of a) Level 4.2.c1 and b) 4.2.c2: a) Cross point (marked by a black circle) and cross (marked as a black cross) of the right cube correspond to each other: In order to solve the cube, the element with the marked red square has to be rotated in the red line of the marked cross. b) Cross point (marked by a black circle) and cross (marked as a black cross) of the right cube do not belong together: In order to solve the cube, the element with marked yellow square has to be rotated on the left side of the back, in the invisible yellow line, which is on the back of the right cube.

Figure 5: Impossible sample items of the six difficulty levels of the R-Cube-Vis Test. Note, the impossible items of two most difficult levels are indistinguishable.

This construction allows the generation of many items of the same kind with respect to structure and difficulty, with the difficulty being dependent on the cubes' size and number of the rotated elements as well as how these elements are rotated. The visual appearance can be varied by manipulating the color arrangement, the axes used, the position of the rotated elements for a specific axis, and the direction of the rotation. The expectation was that the item difficulty is independent of these aspects of visual appearance while the manipulated features of the visual appearance ensure that items appear different from each other. One could therefore expect that all items that were alike with respect to structure and difficulty could be assigned to one homogenous group and that the variability of the visual appearance within these groups guaranteed that no item could be solved by remembering the answer to a previously seen item of the same group.

Due to this rational construction of the stimulus materials, the validity evidence based on test content is given according to the standards formulated by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). The content-oriented evidence "can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure." (American Educational Research Association et al., 2014, p. 14). This evidence can be confirmed for the R-Cube-Vis Test regarding the definition of visualization (see Section 2). In case of the R-Cube-Vis Test, the "image of spatial patterns" (here, the right cube of an item) has to be "manipulated/transformed" (rotating single elements, in several consecutive steps) to receive an "other arrangement" (the left cube).

Test measures

The participants had to provide two answers for each item. First, they had to decide whether both cubes are possibly the same. The resulting accuracy measure was coded with 1 (correct answer) and 0 (incorrect answer). Second, the participants had to estimate their confidence about the given answer on a scale from 1 (very sure) to 4 (very unsure).

Based on these two answers per item, four different accuracy measures were evaluated for the R-Cube-Vis Test. An optimal measure does not exist. The theoretical drawbacks of the four measures will be discussed below. All measures were analyzed with respect to reliability and validity in Study 1 – 3 since all measures differ in the information they use.

The first measure is ACC-all. It is calculated as the sum of the item accuracies over all items divided by the number of all items. The second measure is ACC-poss. It sums the item accuracies of all *possible* items and divides it by the number of all possible items. Both measures can reach values ranging from 0 (all considered items were answered incorrectly) to 1 (all considered items were answered correctly). Whereas ACC-poss focuses exclusively on items that were presumably solved by using the intended transformations, i.e., the rotation of the cube's elements (the possible items), ACC-all uses twice as many items, can control for answering bias and is, therefore, more robust than ACC-poss. For example, if a participant always ticks “is possible”, she would receive an ACC-poss score of 1, which would indicate a high ability, but would receive an ACC-all score of .5, which would correctly indicate a low ability.

The main drawback of both measures is the expected statistical error given 50% chance level. To reduce this error, a weighted accuracy (wACC) was introduced in the present studies. This accuracy measure uses the accuracy of each possible item and the corresponding confidence rating. The wACC is the product of the item accuracy (0, incorrect answer or 1, correct answer) and the reversed confidence rating value (1 to 4). Therefore, the wACC-value of each item ranges from 0 to 4. For example, if a participant answers an item correctly and gives a confidence rating of 3, then the $wACC = 1 * 3 = 3$. If a participant answers an item incorrectly and gives a confidence rating of 2, then the $wACC = 0 * 2 = 0$. In generally, $wACC = 0$ means that the item is not solved correctly independent of the participant's confidence. Values from 1 (strongly unsure) to 4 (strongly sure) thus indicate a correctly solved item with varying confidences. wACC is based on possible items only, such as ACC-poss. The evaluation of the answer by the additional confidence rating is thought to increase the “true” variance. If a participant is guessing an answer correctly with low confidence, this participant will nevertheless get the highest accuracy value (ACC-all and ACC-poss) for this item (i.e., 1) but only a medium wACC value (i.e., 1 or 2). Although there is evidence that self-ratings of cognitive abilities can be valid (see Bratko, Butkovic, Vukasovic, Chamorro-Premuzic, & Stumm, 2012; Jacobs & Roodenburg, 2014), the confidence rating of the wACC measure might be influenced by personal characteristics such as self-esteem which can lead to systematic errors in wACC.

Therefore, a fourth measure was calculated that combines the answers of all items (possible and impossible, correctly and incorrectly answered items) and their confidence ratings according to signal detection theory (Macmillan & Creelman, 2005). The receiver operating curve (ROC) can be computed based on the participant's ratings ranging from 0 (strongly sure that the item is impossible) to 3 (strongly unsure, impossible) and 4 (strongly unsure, possible) to 7 (strongly sure, possible). The ROC relates the true positives to the false positives for each of the eight categories. For each participant, the area under the curve (AUC) was computed based on the ROC. In comparison to the previous measures, the AUC uses confidence ratings and controls for a personal decision threshold but might be biased by including impossible items and weighting wrong answers.

In summary, no measure fulfills all of the following aspects in an optimal way: low chance level, no answering bias, no personal bias, excluding impossible items, and no weighting of wrong answers. Each measure has drawbacks regarding at least one of these aspects. However, the measures ACC-all and ACC-poss are standard measures of current performance tests and, therefore, are appropriate performance measures. The innovative measures that consider confidence ratings additional to accuracy, wACC and AUC, are alternatives that potentially compensate drawbacks of the standard measures, but have their own drawbacks (see above). Given this ambiguousness, all measures will be reported in the validation studies to

determine the most suitable one for the test. All four measures will be analyzed and compared with respect to reliability and validity aspects.

Study 1: Development and validation of a long version of the R-Cube-Vis Test

Study 1 had two goals. First, the results of the preliminary study should be confirmed with respect to the order of the difficulty levels and the correlations with external variables. For all six defined R-Cube-Vis levels (Figure 3), additional items were created. The expectation was that the order of the levels regarding their difficulty is similar to the results found in the preliminary study: Level 4.1, 3.1, 4.2.p, 3.2.p, 4.2.c1, 4.2.c2. Size 4 should be easier than size 3, one rotated element is easier than two, and parallel elements are easier than crossed ones. It is possible that items of size 4 are easier than items of size 3 because more parts of the cubes' surfaces of the items with size 4 stay the same after rotating the elements than of the cubes' surfaces of items with size 3. For example, the right cube of an item of Level 4.1 (size 4, Figure 3) shows on the two sides with the rotated element eight changed pieces and 24 pieces that stayed the same. That means that $8 / (24 + 8) = 1 / 4$ of the pieces changed. Whereas the right cube of an item of Level 3.1 (size 3, Figure 3) shows on the two sides with the rotated element six changed pieces and twelve pieces that stayed the same. Therefore, relatively more items changed, $6 / (12 + 6) = 1 / 3$. A reason for the fact that the items with parallel rotated elements are easier to solve than items with crossed rotated elements might be that the order of the back rotation to solve the item with two parallel rotated elements (Level 4.2.p, Figure 3) is arbitrary, whereas the order of the back rotation of the two crossed rotated elements is fixed (Level 4.2.c1, 4.2.c2, Figure 3).

Furthermore, the test was analyzed with respect to validity *evidence based on relations to other variables* (American Educational Research Association et al., 2014). This validity aspect refers to the relation to “external variables [that] may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs.” (p. 16) In particular, evidence for convergent validity was examined regarding the standard visualization test, the Paper Folding Test (PFT), as well as evidence for discriminant validity with respect to the established spatial relations tests. The correlations with the Mental Rotation Test (MRT-A) should lie in between due to its position in the middle of the continuum between visualization and spatial relations (Section 2). The school grades in German and Mathematics were also considered for comparing their correlations with the validation tests and R-Cube tests.

Second, validity *evidence based on internal structure* (American Educational Research Association et al., 2014) was examined for the R-Cube-Vis Test. This covers the “analyses of the internal structure of a test [that] can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based.” (p. 16). The expectation was that all items of the R-Cube-Vis Test lie on one dimension and that all items of each of the six levels have a unique level of difficulty. Hence, the R-Cube-Vis Test was tested for conformity with the linear logistic test model (LLTM, Fischer, 1973).

In Study 1 – 2, a second developed spatial thinking test, the R-Cube-SR test, was also conducted for validation purposes. This test is intended to measure the second spatial thinking factor, spatial relations (therefore: R-Cube-SR), and uses the same cubes but in a changed presentation. This test will be published in a separate article and is not analyzed in the present paper.

Method

Participants

The minimal sample size to find the smallest expected correlation between the two spatial relations tests, $r = .34$ (see Section 2), with a directed hypothesis, an α -level of .05, and a test power at .80 (recommended by Cohen, 1988) is 49 (G*Power, Version 3.1.9.2, Faul, Erdfelder, Lang, & Buchner, 2007). Study 1 was therefore

conducted with 53 non-colorblind participants (40 female, 13 male) from the University of Mannheim. They received course credit for their participation. The participants' mean age was $M = 21.15$ years ($SD = 2.15$ years), ranging from 17 to 26 years. The final sample consisted of $N = 49$ (39 female, 10 male), because three participants were excluded according to the exclusion criteria defined based on the data of the preliminary study, and one participant because he misunderstood the PFT instruction. Furthermore, speeded tests, such as a chronometric test (here termed "Chronometric Test" and abbreviated "CT", Jansen-Osmann & Heil, 2007, see below), assume that participants can easily solve the tasks, and that they mainly differ in their reaction times and not in their accuracy. Therefore, participants were also excluded from correlations with the CT if they had an error rate higher than 30% on all possible items in order to guarantee that the reaction times reflect the speed of the cognitive processes to solve these items and not just guessing behavior. The same threshold was also used by Jansen, Schmelter, Quaiser-Pohl, Neuburger, and Heil (2013) for another speeded mental rotation test. Hence, the analyzed correlations with the CT are based on only $N = 45$ participants. This study as well as the following studies were administered in compliance with the guidelines of the ethic committee.

Materials

The validation tests were PFT, MRT as well as two spatial relations test, the Würfelaufgaben (form A) from the Intelligenz-Struktur-Test, IST-2000R, abbreviated as IST-WA (Amthauer, Brocke, Liepmann, & Beauducel, 1999) and the CT. In the IST-WA, participants have to decide whether cubes are the same except for the rotation of one cube as a whole (no single elements as in the R-Cube-Vis Test). Items of the CT show two 2D-object pairs ("primary mental ability figures", Thurstone, 1958) that participants have to rate as either same or mirrored except for rotation.

PFT, MRT, and IST-WA were presented as paper-pencil tests, and the CT was computer-based. The PFT was conducted in its long version, with 20 items presented on two sheets of paper with ten items on each sheet. For each page, participants had three minutes to solve the items. The sum score was computed based on all correctly answered items (correct answer: 1 point; incorrect answer: 0 points). The 24 MRT items were also presented on two pages with twelve items on each page. Participants had four minutes to solve the items on each page. For each item, participants received 1 point only if both correct figures were detected, otherwise they received 0 points. In each of the 60 items of the CT, participants had to decide whether both objects are mirrored or equal except for rotation. There were 16 items, eight same and eight mirrored, for 45° , 90° , 180° disparity angles. Twelve additional items had a disparity angle of 0° and are mirrored. Each correctly answered item resulted in 1 point, otherwise 0 points. The CT is constructed as a speed test, and therefore, in all analyses, the logarithmized reaction times of the correctly solved items were used. These reaction times were computed based on the subset of the 24 possible items (all twelve items with 0° disparity angle were impossible) that were answered correctly. The IST-WA was administered with 20 items. Participants received 1 point for each item answered correctly, and 0 points for item answered incorrectly.

The R-Cube-Vis Test was conducted in a long version with 24 possible and 24 impossible items for each level. The test was computer-based, whereby the presentation of the instruction and the repetition of the trial phase were self-administered (i.e., participants decided about their reading times and whether they wanted to repeat practice trials) and the presentation of the items, and the recording of the answers were system-administered (i.e., controlled by the computer). Within each level, items were varied systematically, regarding (1) the three axes about which the elements were rotated, (2) the turned element (front, middle, end) and (3) the turning direction. The R-Cube-Vis Test was conducted in five blocks with one block for each level, except the two most difficult levels. The block order was fixed beginning with the cubes of size 3 followed by the cubes of size 4 (3.1; 3.2.p; 4.1; 4.2.p; 4.2.c1 and 4.2.c2 together). All items within each block were randomly presented. The trial phase before each block consisted of five items, two possible and three impossible. Participants had the opportunity to repeat this phase. Please note that at this stage of the development of the R-Cube-Vis Test, the two most difficult levels (Level 4.2.c1 and 4.2.c2) were not clearly identified as separate levels, because the results of the preliminary study delivered only weak evidence.

Therefore, items of both levels were handled as belonging to the same group and were presented in the same block. However, these items were balanced within the block with 12 possible and 12 impossible items per level. Hence, the R-Cube-Vis Test was conducted with five blocks and 240 items overall.

Items of the R-Cube-SR Test show two Rubik's Cubes, whereby the participants have to decide whether it is possible to rotate the left cube as a whole to get the right cube. The items were provided in two blocks. The cubes of the first block have single-colored sides, whereas the items of the second block show patterns on their sides. Each block contained twelve possible and twelve impossible items. The R-Cube-SR Test was also computer-based like the R-Cube-Vis Test with self-administered (i.e., participants decided about their reading times and whether they wanted to repeat practice trials) as well as system-administered (i.e., controlled by the computer) parts.

Participants also had to fill out a questionnaire asking for age, gender, experience with Rubik's Cube and the Abitur grades in German and Mathematics (Abitur is the diploma from German secondary school qualifying for university admission. The grades range theoretically from 0.7, best grade, to 6.0, worst grade.).

Procedure

There were two sessions, A and B, with durations of 90 and 60 minutes. In session A, participants conducted the PFT followed by the R-Cube-Vis Test. Session B took place one week later. The participants performed the MRT, IST-WA, R-Cube-SR, CR, and the questionnaire in the indicated order. The participants were tested in groups of up to six people. Here, and in all the following studies, all computer-based tests were presented with the presentation software E-Prime 2.0 (Psychology Software Tools Inc., 2012).

Analyses

All analyses in this and the following studies were conducted by R statistics (R Core Team, 2017) using the packages `plyr` (Wickham, 2011), `dplyr` (Wickham, François, Henry, & Müller, 2018), `reshape2` (Wickham, 2007), `zoo` (Zeileis & Grothendieck, 2005) to prepare the data and perform descriptive statistics. Pearson correlations were computed using the package `psych` (Revelle, 2017). Correlation comparisons were tested with Fisher's z by the package `cocor` (Diedenhofen & Musch, 2015). The significance levels of all pairwise correlations of the validation tests, as well as all correlations between each validation test and both grades, in German and Mathematics, were adjusted according to Holm (1979). The same adjustment was also applied to test evidence of convergent and discriminant validity with the validation tests and to test evidence of validity of the external criteria using the grades in German and Mathematics. All adjustments were made separately for each tested measure of the R-Cube-Vis Test. Reliability estimates were computed as standardized Cronbach's α (Cronbach, 1951) using R-packages `psy` (Falissard, 2012) and `psych` (Revelle, 2017). In all analyses, reaction times were logarithmized and only considered for the correctly answered possible items.

In order to examine evidence of the validity of the internal structure, the R-Cube-Vis stimulus materials were tested for conformity with the linear logistic test model (LLTM) to provide evidence for one-dimensionality and homogeneity of the six levels of difficulty, Level 3.1, 3.2.p, 4.1, 4.2.p, 4.2.c1, and 4.2.c2. The focus was on both measures that were defined for all six levels, accuracy over all possible items (ACC-poss), and weighted accuracy (wACC). Note that the impossible items in Level 4.2.c cannot be assigned to one of the two subgroups. Therefore, accuracy over all items (ACC-all) and the Area Under the Curve (AUC) were not considered here. The design matrix of the LLTM consisted of four parameters indicating (1) size 3 or size 4, (2) one or two rotated elements, (3) parallel or crossed rotation, and (4) whether the cross and the corresponding cross point are both visible. Hence, six item categories resulted, one for each level of difficulty. The model conformity was tested according to the procedure described in Baghaei and Kubinger (2015). As a prerequisite, the standard Rasch model (RM) was performed (Fischer, 1973) and tested with Andersen's likelihood ratio test (Andersen, 1973) using mean split of mean raw scores. Additionally, the

item characteristics were analyzed by their infit (weighted mean square). All analyses were conducted using the R-package `eRm` (Mair, Hatzinger, & Maier M. J., 2015). In the last step, both models, LLTM and RM, were compared using a chi-square test given the deviance of their $-2 \cdot \log\text{likelihood}$'s as empirical value and the difference between their number of model parameters as degrees of freedom (Fischer, 1973). Furthermore, the information criteria AIC (Akaike Information Criterion, Sakamoto, Ishiguro, & Kitagawa G., 1986) and BIC (Bayesian Information Criterion, Schwarz, 1978) were compared between the RMs and the corresponding LLTMs. These information criteria indicate a compromise between model fit and simple model structure. The differences between two AIC values (two BIC values), Δ_{AIC} (Δ_{BIC}), were interpreted as negligible for $\Delta_{\text{AIC}} \leq 2$, weak for $2 < \Delta_{\text{AIC}} < 4$ ($\Delta_{\text{BIC}} \leq 2$), moderate for $4 \leq \Delta_{\text{AIC}} \leq 7$ ($2 < \Delta_{\text{BIC}} \leq 6$), strong for $7 < \Delta_{\text{AIC}} \leq 10$ ($6 < \Delta_{\text{BIC}} \leq 10$), and very strong for $\Delta_{\text{AIC}} > 10$ ($\Delta_{\text{BIC}} > 10$) following Burnham and Anderson (2016) for AIC and Raftery (1995) for BIC. For both information criteria, lower values indicate the more appropriate model. For both model types, RM and LLTM, the person separation reliability (PSR) was computed, which, according to Wright and Stone (1999), can be interpreted similarly to the KR20 reliability index (Kuder & Richardson, 1937).

Both models in their standard form assume dichotomous items with expected values ranging from 0 to 1. Therefore, the RM and the corresponding LLTM could not be directly applied to ACC-poss because of the guessing probability at 50% and to wACC because of the five answer categories. In case of ACC-poss, additionally to a simple RM, a 3-PL model (Birnbaum, 1968) with a fixed guessing parameter at .5 and a fixed discrimination parameter at 1 was adapted to the data. The 3-PL model was compared with the RM using the `ltm`-package (Rizopoulos, 2006) with respect to the information criteria AIC and BIC. The corresponding LLTM of the most appropriate model was tested in the following way. The values of wACC were dichotomized with splits between the values 1 and 2, 2 and 3, and 3 and 4. These splits resulted in three new dichotomous variables indicating 0 for all values lower than the splitting threshold and 1 for all values larger than the splitting threshold, i.e., for a split between 2 and 3, the wACC values 0, 1, and 2 result in the new value 0, whereas the values 3 and 4 result in the new value 1. For each of these splits, RMs and LLTMs were conducted and tested as described above. Note that a split between 0 and 1 would result in a new variable identical to ACC-poss.

Two issues arose due to the small sample size. In a case where all participants have the same value for some items, the estimated RMs and LLTMs, as well as the models that were estimated for the Andersen's test, might be based only on a subsample of items. In this case, all items with only one value in the respective sample had to be excluded. Furthermore, the statistical power is low for both likelihood ratio tests. However, robust parameter estimations are still expected (see Hohensinn, Kubinger, & Reif, 2014; MacDonald & Kromrey, 2011). In addition to the statistical tests, the results were visually inspected using the R-package `ggplot2` (Wickham, 2009).

Results

Descriptive results (Table 1) show the not-logarithmized reaction times of the CT. However, according to the results of the preliminary study that showed a non-normal distribution of the reaction times, the logarithmized reaction times of the CT were used for further analyses. Although some measures were significantly different from the normal distribution according to the Shapiro-Wilk test of normality (Royston, 1995), Pearson's correlations as a robust method (e.g., Havlicek & Peterson, 1976) were computed in subsequent analyses.

The inter-correlations between the validation tests of both spatial thinking factors, PFT, MRT, and IST-WA were similar, $.49 \leq r \leq .55$, $p < .01$ (Table 2). Only the correlations with CT showed the expected trend with $r \geq -.37$, $p < .05$ for PFT and MRT, and with a higher correlation with the second spatial relations test, IST-WA, $r = -.56$, $p < .001$. However, neither the correlation between PFT and MRT nor the correlation between both spatial relations tests were significantly different from inter-correlations between both spatial thinking factors, $z \leq 1.29$, $p \geq .20$.

Table 1: Descriptive results of each test

	N	M	SD	Range		Skewness	Kurtosis
PFT	52	.65	.16	.25 -	.95	-.50	-.24
MRT	53	.60	.21	.08 -	.96	-.37	-.91
IST-WA	53	.59	.23	.15 -	1.00	.08	-.88
CT (ACC)	49	.92	.07	.71 -	1.00	-1.05	1.11
CT (RT)	49	2454.12	606.76	1291.78 -	3792.54	.44	-.69
R-Cube-Vis (ACC-all)	50	.88	.07	.65 -	.98	-.94	.83
R-Cube-Vis (ACC-poss)	50	.90	.07	.66 -	.99	-1.38	2.27
R-Cube-Vis (wACC)	50	3.25	.41	2.33 -	3.86	-.47	-.88
R-Cube-Vis (AUC)	50	.96	.05	.77 -	1.00	-2.15	4.92

Note. The different sample sizes result from different excluded participants for each test.

Table 2: Correlations between visualization and spatial relations tests

	PFT	MRT	IST-WA
MRT	.49**		
IST-WA	.52**	.55**	
CT	-.37*	-.34*	-.56***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. All correlations with CT are based on $N = 45$ participants.

The grade in Mathematics showed descriptively higher but not significant correlations with the validation tests ($.26 \geq |r| \geq .30$, $p \geq .08$) compared to the grade in German, ($.06 \geq |r| \geq .16$, $p \geq .28$). The highest correlations with experience with Rubik's cubes were found for PFT, $r = .32$, $p < .05$, and IST-WA, $r = .25$, $p = .08$. MRT and CT were less correlated, $|r| \leq .11$, $p \geq .43$.

R-Cube-Vis Test for visualization: Descriptive results

The five R-Cube-Vis levels showed the expected order of increasing difficulty over the defined levels regarding all four measures (accuracy based on all items, ACC-all, accuracy based on all possible items, ACC-poss, weighted accuracy, wACC, Area Under the Curve, AUC, Table 3), although Level 4.1 and 3.1 had nearly the same values (Table 3).

Table 3: Means (standard deviations) of the four R-Cube-Vis Test measures: ACC-all, ACC-poss, wACC, and AUC, also for all levels separately

	ACC-all	ACC-poss	wACC	AUC
All levels	.88 (.07)	.90 (.07)	3.25 (0.41)	.96 (.05)
Level 4.1	.97 (.05)	.99 (.02)	3.80 (0.31)	.97 (.04)
Level 3.1	.96 (.06)	.98 (.04)	3.68 (0.41)	.95 (.05)
Level 4.2.p	.90 (.11)	.96 (.10)	3.53 (0.57)	.94 (.05)
Level 3.2.p	.86 (.12)	.92 (.13)	3.24 (0.61)	.92 (.05)
Level 4.2.c	.69 (.13)	.64 (.21)	1.93 (0.80)	.90 (.05)
Level 4.2.c1	-	.75 (.24)	2.42 (1.05)	-
Level 4.2.c2	-	.53 (.29)	1.41 (0.96)	-

The two levels Level 4.2.c1 and 4.2.c2 showed the expected differences with respect to ACC-poss (.75 for 4.2.c1; .53 for 4.2.c2) and wACC (2.42; 1.41). ACC-all and AUC were not computed for both levels because the impossible items cannot be unambiguously assigned to one of these subgroups.

Table 4: Correlations of the R-Cube-Vis Test and its single levels with the validation tests

	Measure	PFT	MRT	IST-WA	CT
All levels	<i>ACC-all</i>	.58***	.34	.31	-.26
	<i>ACC-poss</i>	.51***	.27	.18	-.02
	<i>wACC</i>	.52***	.41*	.33*	-.22
	<i>AUC</i>	.59***	.34*	.32*	-.17
Level 4.1	<i>ACC-all</i>	.33	.17	.20	-.20
	<i>ACC-poss</i>	.29	.09	.22	-.11
	<i>wACC</i>	.16	.10	.18	-.03
	<i>AUC</i>	.25	.29	.25	-.06
Level 3.1	<i>ACC-all</i>	.39*	.32	.27	-.21
	<i>ACC-poss</i>	.35	.22	.17	.09
	<i>wACC</i>	.33	.34	.35	-.15
	<i>AUC</i>	.41**	.56***	.40**	-.25
Level 4.2.p	<i>ACC-all</i>	.40*	.21	.20	-.20
	<i>ACC-poss</i>	.42**	.22	.23	.00
	<i>wACC</i>	.47**	.39*	.32*	-.13
	<i>AUC</i>	.63***	.53***	.49***	-.34*
Level 3.2.p	<i>ACC-all</i>	.34	.12	.11	-.04
	<i>ACC-poss</i>	.26	.09	.05	-.01
	<i>wACC</i>	.36*	.31	.23	-.13
	<i>AUC</i>	.45**	.43**	.43**	-.23
Level 4.2.c	<i>ACC-all</i>	.67***	.43**	.43**	-.38*
	<i>ACC-poss</i>	.39*	.23	.13	-.06
	<i>wACC</i>	.54***	.32	.24	-.32
	<i>AUC</i>	.58***	.51***	.40**	-.37*
Level 4.2.c1	<i>ACC-poss</i>	.42**	.46**	.36*	-.10
	<i>wACC</i>	.50***	.43**	.41**	-.30*
Level 4.2.c2	<i>ACC-poss</i>	.21	-.06	-.12	.00
	<i>wACC</i>	.35	.05	-.06	-.20

Note. * $p < .05$, ** $p < .01$, *** $p < .001$ (significance level was adjusted for each R-Cube-Vis measure and each level separately). All correlations with $|r| \geq .3$ are bold.

Reliability and evidence for convergent and discriminant validity

The reliability estimates for all measures of the R-Cube-Vis Test were very good, $\alpha = .94$ (ACC-all), $\alpha = .90$ (ACC-poss), $\alpha = .97$ (wACC). All R-Cube-Vis measures with the PFT were highly correlated, $.51 \leq r \leq .59$, $p < .001$. The correlations with all other validation tests (MRT, IST-WA, CT) decreased descriptively in that order

for all measures (Table 4). However, the differences between the correlations were only significant between PFT and CT and only for ACC-poss and AUC, $z \geq 2.36$, $p < .05$.

The correlations with the grade in German were negligible, $|r| \leq .06$, $p \geq .71$. The strongest correlation with the grade in Mathematics was $r = -.20$, $p \geq .17$ for ACC-all, ACC-poss, and AUC. wACC was less correlated with the grade in Mathematics, $r = -.12$, $p = .40$. Correlations with experience with Rubik's cubes were not significant and ranged from $r = .06$, $p = .66$ (ACC-all) to $r = .24$, $p = .09$ (wACC).

Testing conformity with the linear logistic test model (LLTM)

In a first step, the items were analyzed to test the Rasch model (RM) against the 3-PL model for ACC-poss as well as the hypothetical assumptions against polytomous IRT-models using the wACC (see Section 4.1.4).

The first analysis showed strong evidence in favor for the RM compared to the 3-PL model for ACC-poss according to AIC and BIC (RM: AIC = 2714.33, BIC = 2895.98; 3-PL: AIC = 2725.01, BIC = 2906.65). Therefore, the RM and its corresponding LLTM were for ACC-poss and the dichotomized wACC described above (Section 4.1.4) in the following analyses.

Results of Andersen's likelihood ratio test of RMs showed good fits for ACC-poss, and all splits of wACC, $p \geq .98$ (Table 5), which is supported by the plots showing the estimated β 's for low and high ability groups pitted against each other (Figure 6). All reliability estimates were good, $PSR \geq .84$, and at least 99% of all infits were in the acceptable range from 0.5 to 1.5 according to Linacre (2002, Table 5).

Table 5: Results of Andersen's likelihood ratio test, infits and reliability estimates (PSR) of the RMs

	χ^2	df	p	Infits between 0.5 and 1.5 [%]	PSR (RM)
ACC-poss	37.45	58	.98	100	.84
Between 1 and 2	28.64	56	1.00	100	.88
Between 2 and 3	47.35	76	1.00	100	.92
Between 3 and 4	72.21	102	.99	99	.98

Table 6: Comparison between RMs and corresponding LLTMs using log-likelihoods and reliability estimates (PSR) for LLTMs

	χ^2	df	p	PSR (LLTM)
ACC-poss	62.22	90	.99	.84
Between 1 and 2	78.78	94	.87	.88
Between 2 and 3	93.64	105	.78	.92
Between 3 and 4	111.86	115	.57	.98

The corresponding LLTM's were not significantly different from the RMs, $p \geq .57$ (Table 6) and showed similar β estimations (Figure 7). The reliability estimates were good, $PSR \geq .84$ (Table 6) and equal to the estimates of the corresponding RMs (Table 5). For all splits of wACC as well as ACC-poss AIC and BIC showed strong evidence for the LLTMs compared to the RMs with $\Delta AIC > 100$ and $\Delta BIC > 100$.

Figure 6: Normalized β -parameters for RMs estimated from both subsamples used for Andersen's likelihood ratio test, below and above the mean.

Figure 7: Comparison between normalized β -parameters for RMs and corresponding LLTMs.

Discussion

The first goal of Study 1 was the validation of the long version of the R-Cube-Vis Test with respect to convergent and discriminant evidence of validity. Secondly, the internal structure of the R-Cube-Vis Test was analyzed, and it was tested whether the six levels of difficulty define homogenous item groups.

All correlations between these validation tests were in the expected ranges known from the literature (e.g., Blajenkova et al., 2006; Kozhevnikov et al., 2002; Kozhevnikov & Hegarty, 2001; Miyake et al., 2001, see also Section 2). However, the correlation between PFT and MRT was descriptively smaller than the correlations of both tests with the IST-WA. The correlations with the CT descriptively showed the expected pattern with a stronger correlation with other spatial relations test, IST-WA and weaker correlations with the PFT and MRT. The correlations with the grades were negligible for the German grade but reached a medium effect size for the grade in Mathematics ($r \approx .30$). Both results are consistent with the current literature (e.g., Geary et al., 2000; Gunderson et al., 2012; Kaufman, 2007; LeFevre et al., 2010, see also Section 2).

The descriptive results of the R-Cube-Vis Test confirmed the order of the six levels with respect to their difficulty for all four measures. The reliability of the complete test was very good concerning the accuracy over all items (ACC-all), the accuracy over all possible items (ACC-poss), and the weighted accuracy (wACC).

Correlations between the R-Cube-Vis Test and all four validation tests showed convergent and discriminant evidence of validity through strong correlations with the PFT, weaker correlations with the MRT and the weakest correlations with both spatial relations tests, IST-WA and CT. A small difference between the correlations with IST-WA and CT might be based on the different measures, as IST-WA uses an accuracy-based measure as do all measures of the R-Cube-Vis, whereas the CT uses reaction times. Similar correlation patterns with the validation tests were also found for single levels. This supports the conclusion that the test materials are homogeneous. The smaller correlations of Level 4.1 and 4.2.c2 can be explained by ceiling and floor effects, respectively. All four measures are suitable for the R-Cube-Vis Test as a visualization test. Evidence of validity of external criteria, the grade in German and Mathematics, was weak but comparable to the correlations between these grades and the validation tests.

Furthermore, evidence of validity of the internal structure was shown by the confirmation of the R-Cube-Vis Test with the linear logistic test model (LLTM). Although the small sample size is limiting the interpretation of both likelihood tests due of the reduced power, one can reasonably expect the parameters to have been accurately estimated (see Hohensinn et al., 2014; MacDonald & Kromrey, 2011). Therefore, the main focus is on further criteria to test the model conformity. Infit ranges, person separation reliability (PSR), as well as the visual inspection of the beta-value-correspondence plots, support the conformity with the RM. The nearly equal PSR values of the corresponding LLTMs, as well as the lower information criteria for the LLTMs for ACC-poss and all splits of wACC, show that the LLTMs are more appropriate models for the R-Cube-Vis Test than the RMs.

The results support the R-Cube-Vis Test as a one-dimensional visualization test with six discrete difficulty levels for all four tested measures.

Study 2: Development and validation of short versions of the R-Cube-Vis Test

The complete R-Cube-Vis Test from Study 1 took up to 50 min for all 240 items. The goal of Study 2 was the development and validation of a short version of the test. To this end, 10,000 item samples, five possible items from each of the difficulty levels (30 possible items in total), were randomly drawn from the complete test materials created for Study 1. For each item sample and based on the data of Study 1, the weighted accuracy (wACC) was computed and correlated with the Paper Folding Test (PFT). The two item samples with the strongest ($r = .66$, $p < .001$), and weakest ($r = .40$, $p < .01$), correlation were chosen as short versions of the R-Cube-Vis test for the present study. The expectation was that both samples would show similar results regarding reliability and evidence of validity because all items from a single level should be equivalent. The strong and weak correlation with the PFT should be due only to statistical error and not result from

systematic differences between the single items. A partial confirmation of this assumption is suggested by the fact that five of the 30 items were identical in both item samples. Both item samples were analyzed with respect to reliability and to convergent, discriminant, criterion, and construct evidence of validity.

Method

Study design

The between-subject design of the present study consisted of two conditions, A and B. In condition A, the participants performed the short version with the item set of the R-Cube-Vis Test that was strongly correlated with the PFT, whereas in condition B, the participants performed the short version with the item set that was weakly correlated with the PFT. All other conducted tests were identical in both conditions.

Participants

The correlations found in Study 1 were slightly weaker than expected from the literature. The sample size was, therefore, increased in Study 2 and 3. The sample consisted of $N = 119$ participants, 99 female and 20 male, which were from 18 to 39 years old ($M = 22.29$ years, $SD = 3.08$ years). They were all students at the University of Mannheim, they were not color-blind, and they could choose between course credit and monetary compensation for their participation. In condition A, $N = 59$ (49 female, 10 male) participants and in condition B, $N = 60$ (50 female, 10 male) participants took part. The same exclusion criteria as in the preliminary study and Study 1 were applied in Study 2. The final subsamples of conditions A and B consisted of $N = 55$ (46 female, 9 male) and $N = 53$ (44 female, 9 male) participants to test evidence aspects of the validity of the R-Cube-Vis Test.

Materials

The tests PFT, Mental Rotation Test (MRT), and CT as well as the questionnaire were the same as in Study 1. All tests were conducted and assessed as described before (Section 4.1.2). Instead of the IST-WA, the Cube Comparison test (CC, Ekstrom et al., 1976a, Figure 8) was administered as second spatial relations test. The CC also utilizes cubes as stimulus materials with letters and symbols on their sides, whereby all sides of one cube are pairwise different. Each item shows two cubes. Participants have to decide whether both cubes are the same except for rotation (“same”) or not (“different”).

Figure 8: Cube Comparison Test, CC. Here, both cubes are different.

In accordance with Ekstrom et al. (1976b), participants had 3 minutes for each of the two test parts of the CC. Each part contained 21 items. One item could be interpreted ambiguously as “same” or “different”, because one cube side shows an “X” and it is very hard to see if the “X” on the right cube is lying on the side

or standing upright. Therefore, it was excluded from the following analyses. The total number of items was 41. The participants received one point for each correctly answered item of the CC and 0 points otherwise.

There were two short versions of the R-Cube-Vis Test, presented in condition A and B, as described above. Each short version consisted of six levels, and each level contained five possible and five impossible items. Each of the R-Cube-Vis short versions consisted of six blocks, one for each of the six levels. The blocks were presented in order of increasing difficulty in accordance with the results of Study 1 (4.1; 3.1; 4.2.p; 3.2.p; 4.2.c1; 4.2.c2). At the beginning of the test, there was a trial phase containing six items, one from each level (three possible and three impossible). Participants were given the opportunity to repeat the trial phase. Same as in Study 1, the test was computer-based and self- as well as system-administered.

Procedure

In both conditions, participants conducted the three paper-pencil tests, PFT, MRT and CC followed by the CT. Afterwards, they performed one of the two R-Cube-Vis short versions in either condition A or condition B. Finally, the participants performed the R-Cube-SR and answered the descriptive data questionnaire. The entire session took 90 minutes to complete.

Results

The accuracy of all tests was slightly lower than in Study 1 (Table 7). The results of all measures of the R-Cube-Vis Test in both conditions were comparable. As in Study 1, for analyses with the CT, participants were excluded if they had an error rate higher than 30% on all possible items. All validation tests were strongly and similarly correlated with each other ($.50 \leq |r| \leq .57$, $p < .001$) and pairwise not significantly different, $z \geq .45$, $p \geq .65$ (Table 8).

Table 7: Descriptive results of each test

	N	M	SD	Range	Skewness	Kurtosis
PFT	119	.64	.18	.20 -1.00	-.27	-.49
MRT	119	.52	.20	.08 - .96	-.04	-.51
CC	119	.66	.18	.17 -1.00	-.42	-.38
CT (ACC)	92	.90	.08	.71 -1.00	-.72	-.48
CT (RT)	92	3073.33	1218.06	1427.10 -8506.71	1.83	4.61
Condition A						
R-Cube-Vis (ACC-all)	55	.82	.09	.60 - 0.98	-.32	-.81
R-Cube-Vis (ACC-poss)	55	.82	.11	.47 - 1.00	-.39	.07
R-Cube-Vis (wACC)	55	2.76	.50	1.73 - 3.73	-.19	-.53
R-Cube-Vis (AUC)	55	.93	.06	.77 - 1.00	-1.02	.20
Condition B						
R-Cube-Vis (ACC-all)	53	.83	.09	.63 -1.00	-.21	-.71
R-Cube-Vis (ACC-poss)	53	.81	.10	.57 -1.00	-.03	-.24
R-Cube-Vis (wACC)	53	2.74	.57	1.67 -3.93	-.07	-.77
R-Cube-Vis (AUC)	53	.93	.06	.74 -1.00	-1.06	.47

Note. The different sample sizes result from different excluded participants for each test. The non-logarithmized reaction times are presented here.

Table 8: Correlations between visualization and spatial relations tests

	PFT	MRT	CC
MRT	.54***		
CC	.50***	.52***	
CT	-.56***	.57***	-.55***

Note. *** $p < .001$. All correlations with CT based on $N = 86$ participants.

All correlations between the validation tests and the grade in German were negligible and not significant, $|r| \leq .08$, $p \geq .40$. The grade in Mathematics showed stronger (marginal) significant correlations, $r = -.39$, $p < .001$ (PFT), $r = -.17$, $p = .18$ (MRT), $r = -.23$, $p < .05$ (CC), $r = .31$, $p < .01$ (CT). Experience with Rubik's cubes was not significantly correlated with any validation test, $|r| \leq .09$, $p \geq .36$.

R-Cube-Vis Test for visualization: Descriptive results

Descriptive results for R-Cube-Vis stimulus materials were similar for both conditions (Table 9). Compared to Study 1, the order of Level 4.1 and 3.1 was reversed in both conditions regarding their difficulty according to all measures (accuracy based on all items, ACC-all; accuracy, based on all possible items, ACC-poss; weighted accuracy, wACC; Area Under the Curve, AUC). However, the differences between both levels were quite small, as was the case in Study 1. Furthermore, Level 4.2.c1 and 4.2.c2 were solved correctly less often in both conditions than in the previous study.

Reliability and convergent and discriminant evidence of validity

Cronbach's α was comparable in both conditions. Cronbach's α was low for ACC-poss, $\alpha = .69$ in condition A and $\alpha = .67$ in condition B, acceptable for ACC-all, $\alpha = .78/.79$, and good for wACC, $\alpha = .85/.90$.

The correlations with the validation tests showed different results between both conditions (Table 10). In condition B, correlations with the PFT were significant for all measures, $.43 \leq r \leq .49$, $p < .01$. The correlations with the spatial relations tests and the MRT were smaller for ten out of twelve correlations ($|r| \leq .31$) and only two correlations with the spatial relations tests were significant, namely between ACC-all and CC ($r = .37$, $p < .05$) and between wACC and CT ($r = .43$, $p < .05$). In condition A, the correlation pattern over all validation tests differed between the four measures. ACC-all was comparably medium strength but not significantly correlated with all validation tests, whereas ACC-poss showed the strongest correlation with CT, $r = -.47$, $p < .001$. wACC and AUC were most strongly correlated with the PFT ($r = .49$, $p < .001$ and $r = .35$, $p < .05$) and weakly correlated with all other tests.

The grade in German was not correlated with all R-Cube-Vis measures in both conditions, $|r| \leq .09$, $p \geq .51$ (condition A) and $|r| \leq .05$, $p \geq .71$ (condition B). Significant correlations with the grade in Mathematics could be found for ACC-all (condition A/B: $r = -.35/-.36$, $p < .05$, both), AUC ($r = -.41/-.37$, $p < .01/.05$) and for wACC in condition A ($r = -.30$, $p < .05$). All other correlations were negative and not significant, $r \geq -.20$, $p \geq .33$. Experience with Rubik's cubes showed (marginally) significant correlations with ACC-all in condition A only, $r = .32$, $p < .05$, wACC, $r = .28$, $p < .05$, and AUC, $r = .25$, $p = .07$. All R-Cube-Vis measures in condition B were low correlated with experience with Rubik's cube, $|r| \leq .10$, $p \geq .49$.

Table 9: Means (standard deviations) for the four R-Cube-Vis Test measures, ACC-all, ACC-poss, wACC and AUC, also for all levels and both conditions, A and B, separately

	ACC-all		ACC-poss		wACC		AUC	
	A	B	A	B	A	B	A	B
All levels	.82 (.09)	.83 (.09)	.82 (.11)	.81 (.10)	2.76 (0.50)	2.74 (0.57)	.93 (.06)	.93 (.06)
Level 4.1	.95 (.09)	.95 (.11)	.97 (.07)	.98 (.07)	3.65 (0.54)	3.62 (0.59)	.96 (.04)	.95 (.05)
Level 3.1	.97 (.05)	.97 (.07)	1.00 (.03)	.99 (.04)	3.79 (0.37)	3.76 (0.42)	.95 (.04)	.94 (.05)
Level 4.2.p	.87 (.14)	.88 (.10)	.92 (.16)	.94 (.14)	3.12 (0.80)	3.20 (0.86)	.91 (.05)	.93 (.05)
Level 3.2.p	.82 (.15)	.83 (.17)	.88 (.20)	.85 (.22)	2.83 (0.90)	2.81 (1.00)	.91 (.05)	.92 (.05)
Level 4.2.c1	-	-	.66 (.29)	.69 (.27)	1.95 (1.06)	2.01 (1.11)	-	-
Level 4.2.c2	-	-	.49 (.36)	.43 (.29)	1.19 (1.00)	1.06 (0.96)	-	-

Table 10: Correlations of the R-Cube-Vis Test with visualization and spatial relations tests for conditions A and B

	Measure	PFT	MRT	CC	CT (RT)
Condition A	<i>ACC-all</i>	.32	.26	.32	-.33
	<i>ACC-poss</i>	.26	.29	.21	-.47**
	<i>wACC</i>	.49***	.36*	.21	-.36*
	<i>AUC</i>	.35*	.22	.26	-.23
Condition B	<i>ACC-all</i>	.49***	.20	.26	-.31
	<i>ACC-poss</i>	.45**	.26	.37*	-.31
	<i>wACC</i>	.43**	.28	.24	-.43*
	<i>AUC</i>	.46**	.14	.18	-.12

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, significance level was adjusted for each R-Cube-Vis measure and each condition are displayed separately. All correlations with CT (RT) based on $N = 44$ (condition A) or $N = 42$ (condition B) participants. All correlations $|r| \geq .30$ are bold.

Testing conformity with the linear logistic test model (LLTM)

At first, the Rasch model (RM) of ACC-poss was compared with the 3-PL model with a fixed guessing parameter at .5 and a fixed discrimination parameter at 1. As in Study 1, the RM was more suitable with strong evidence according to AIC and BIC in both conditions (condition A: RM: AIC = 1136.17, BIC = 1186.35; 3-PL: AIC = 1168.87, BIC = 1219.05; condition B: RM: AIC = 1090.46, BIC = 1143.66; 3-PL: AIC = 1143.66, BIC = 1166.23). Therefore, the RM was applied for ACC-poss in the following analyses.

The RMs were computed for both conditions separately and tested according to the procedure described in Section 4.1.4 (Table 11, Table 13). For both conditions and all variables (three splits of wACC and ACC-poss), Andersen's likelihood tests showed no significant difference between both ability groups, $p \geq .08$ (Table 11). 100% of all infits were between 0.5 and 1.5. The only exception is the split between 3 and 4 in condition B, with 3% of the infits outside of this range (Table 13). All estimates of the person separation reliability (PSR) were acceptable to very good, $.74 \leq r \leq .94$ (Table 11, Table 13). β -estimates were comparable for the low and the high ability group in both conditions (Figure 9, Figure 11).

Table 11: Results of Andersen's likelihood ratio test, infits and reliability estimates of the three RMs, one for each split for condition A

	χ^2	df	p	Infits between 0.5 and 1.5 [%]	PSR (RM)
ACC-poss	26.87	18	.08	100	.74
Between 1 and 2	18.99	21	.59	100	.77
Between 2 and 3	14.77	21	.83	100	.83
Between 3 and 4	25.80	24	.36	100	.92

Table 12: Comparison between RMs and corresponding LLTMs using log-likelihoods and reliability estimates for LLTMs for condition A

	χ^2	df	p	PSR (LLTM)
ACC-poss	20.23	20	.44	.74
Between 1 and 2	22.45	22	.43	.77
Between 2 and 3	34.87	24	.07	.83
Between 3 and 4	25.87	25	.41	.92

Figure 9: Normalized β -parameters for RMs estimated from both subsamples used for Andersen's likelihood ratio test, below and above the mean, for condition A.

Figure 10: Comparison between normalized β -parameters for RMs and corresponding LLTMs for condition A.

Comparisons between RMs and linear logistic test models (LLTMs) showed no significant difference in the likelihood ratio tests (Table 12, Table 14). The information criteria AIC and BIC showed strong evidence in favor of the LLTMs compared to the corresponding RMs with $\Delta\text{AIC} > 10$ and $\Delta\text{BIC} > 50$ for both conditions. Reliability estimates (PSR) of the LLTMs were nearly the same as for the RMs, $.74 \leq r \leq .93$ (Table 12, Table 14). β -estimates were comparable between all RMs and LLTMs (Figure 10, Figure 12).

Table 13: Results of Andersen's likelihood ratio test, infits and reliability estimates of the three RMs, one for each split for condition B

	χ^2	df	p	Infits between .5 and 1.5 [%]	PSR (RM)
ACC_poss	20.43	18	.31	100	.75
Between 1 and 2	18.83	23	.71	100	.82
Between 2 and 3	27.67	24	.27	100	.87
Between 3 and 4	14.43	21	.85	.97	.94

Table 14: Comparison between RMs and corresponding LLTMs using log-likelihoods and reliability estimates for LLTMs for condition B

	χ^2	df	p	PSR (LLTM)
ACC-poss	12.72	22	.94	.74
Between 1 and 2	23.09	22	.40	.81
Between 2 and 3	27.00	24	.30	.87
Between 3 and 4	32.03	25	.16	.93

Figure 11: Normalized β -parameters for RMs estimated from both subsamples used for Andersen's likelihood ratio test, below and above the mean, for condition B.

Figure 12: Comparison between normalized β -parameters for RMs and corresponding LLTMs for condition B.

Discussion

The goal of Study 2 was the validation of a short version of the R-Cube-Vis Test with two different item samples. These samples were drawn based on the data of Study 1 and selected with respect to their correlations with the Paper Folding Test (PFT). The two item samples with the lowest and highest correlation were chosen to show the interchangeability of the items within a single difficulty level.

All correlations between these tests were similarly strong and lay in the expected interval known from the literature (e.g., Blajenkova et al., 2006; Kozhevnikov et al., 2002; Kozhevnikov & Hegarty, 2001; Miyake et al., 2001, see also Section 2). However, the correlations between the same spatial thinking factor (Visualization, Spatial Relations) were comparable to those between the two different spatial thinking factors. Such similar correlation patterns were found in previous studies as well (e.g., Kozhevnikov et al., 2002; Kozhevnikov & Hegarty, 2001). Similar to Study 1, there were medium effect sizes regarding the correlation with the grade in Mathematics and low correlations between the validation tests and the grade in German.

The reliability estimates of ACC-poss and ACC-all of both R-Cube-Vis short versions were poor to acceptable. However, the true reliability of ACC-all and ACC-poss might be underestimated due to small number of items and their binary format (Brogden, 1946a, 1946b; Sun et al., 2007). wACC showed good reliability estimates in both conditions.

The patterns of the correlations between the R-Cube-Vis short versions and the visualization tests showed different results depending on the conditions. In condition B, all four measures were strongly correlated with the PFT and reached the correlation between the PFT and MRT. This means that the expected correlation pattern could partly be confirmed with strong correlations with the PFT and weaker correlations with the MRT and both spatial relation tests. The results support convergent evidence (for the PFT) and discriminant evidence (for both spatial relations test) of validity. A correlation with the MRT that lies in-between could not be shown. In condition A, the correlation pattern differs between the four measures. The expected pattern could only be found for weighted accuracy, wACC, and Area Under the Curve, AUC. But also in this case, a stronger correlation with the MRT compared to the correlations with both spatial relations tests could not be found. Therefore, a moderate convergent and discriminant evidence of validity only existed for two measures.

Evidence of validity of external criteria, grades in German and Mathematics, were comparable to the correlations found for the PFT with negligible correlations with the grade in German but medium effects for the grade in Mathematics. However, these could only be found for some measures.

The generally weaker correlations of the R-Cube-Vis Test with the PFT and the MRT compared to the reported correlations in Study 1 might result from the changed procedure in Study 2. There was only one trial phase at the beginning with six items, one from each level, either possible or impossible, but not both. Furthermore, the levels were presented in six blocks that were ordered according to increasing difficulty. Due to the smaller number of items within each block, there could be learning effects that might have transferred from one block to the following block, especially if the items were quite similar (e.g., as between Level 4.1 and 3.1). Items from blocks presented later might be easier to solve than items from earlier blocks, as could be seen in the descriptive data (Table 9), where the order between Level 4.1 and 3.1 switched. The procedure was therefore changed to be more appropriate for the short version and tested in the following study.

For both item samples, the Rasch model (RM) conformity was shown for all splits of wACC and ACC-poss regarding all criteria, Andersen's likelihood test, infits and person separation reliability (PSR) as well as visual inspection of the β -estimation plots. Results from the Andersen's likelihood test can only be interpreted carefully due to the small sample size and the resulting low power. Comparisons between the RMs and the corresponding linear logistic test models (LLTMs) confirm the LLTMs as the more suitable models for all splits of wACC and ACC-poss in both conditions. These results support the evidence of validity of the internal structure.

Remarkably, the results support the idea that items within each of the six levels are interchangeable. The order of the correlations with the PFT of both short versions found in Study 1 could not be found in Study 2, it even changed. The item sample of the short version used in condition B was weaker correlated

with the PFT in Study 1, but was stronger correlated with the PFT in Study 2 compared to the short version used in condition A.

Study 3: Validation of an alternative short version of the R-Cube-Vis Test

The goal of Study 3 was to examine a short version of the R-Cube-Vis Test with an alternative procedure to overcome the limited convergent and discriminant evidence of validity found in Study 2. The trial phase was extended to reduce learning effects. Furthermore, items were presented randomly in three combined blocks, where each block consisted of two neighboring levels of difficulty. This revised short version was analyzed with respect to reliability and evidence of validity.

Method

Participants

All $N = 57$ participants, 47 female and 10 male, were students from the University of Mannheim and were not color-blind. On average, the participants were 21 years of age ($M=21.49$, $SD = 3.06$), with the youngest being 18 and the oldest 36. They received course credit for their participation. Participants were excluded according to the same criteria as in all previous studies. The final sample consisted of $N = 52$ (43 female, 9 male) participants.

Materials

The materials were similar to the materials of Study 3, the Paper Folding Test (PFT), the Mental Rotation Test (MRT) and two spatial relations tests, the Cube Comparison Test (CC) and the Chronometric Test (CT). All tests were conducted and evaluated as described in Study 3 (Section 5.1.3).

Participants also had to perform the R-Cube-Vis short version, using the same items from condition B in Study 2. The test consisted of three blocks. The first block contained items from Level 4.1 and 3.1, the second block from Level 4.2.p and 3.2.p and the third block from Level 4.2.c1 and 4.2.c2. All items within each block were presented randomly with five possible and five impossible items from each level. Before each block, the participants had to conduct a trial phase. Each trial phase consisted of four items, one possible and one impossible item from each of the two levels of this block. Participants were given the opportunity to repeat the trial phases. As in the previous studies, the test was computer-based with self-administered and system-administered parts.

Furthermore, a questionnaire was administered for descriptive data, such as age, sex, R-Cube experience, and Abitur grades in Mathematics and German.

Procedure

Participants performed the four validation tests in the order described in Section 6.1.2 followed by the R-Cube-Vis Test and the questionnaire.

Results

Descriptive results (Table 15) showed similar accuracy values compared to Study 2. The correlation between the PFT and MRT was $r = .33$, $p < .05$ and lower than the correlations between the PFT and both spatial relations tests, CC and CT (Table 16). However, both differences did not reach statistical significance, $z \leq -.049$, $p \geq .31$. Both spatial relations tests, CC and CT, were correlated with $r = -.45$, $p < .01$ (Table 16).

Table 15: Descriptive results of each test

	N	M	SD	Range	Skewness	Kurtosis
PFT	57	.65	.19	.20 - 1.00	-.57	-.43
MRT	57	.59	.22	.12 - 1.00	-.24	-1.02
CC	57	.65	.17	.20 - .98	-.41	-.33
CT (ACC)	47	.92	.07	.71 - 1.00	-1.25	1.18
CT (RT)	47	3092.11	964.98	1525.17 - 5244.77	.44	-.42
R-Cube-Vis (ACC-all)	52	.83	.08	.63 - .98	-.34	-.73
R-Cube-Vis (ACC-poss)	52	.83	.09	.50 - .97	-1.02	2.14
R-Cube-Vis (wACC)	52	2.73	.52	1.43 - 3.8	-.33	-.38
R-Cube-Vis (AUC)	52	.93	.06	.72 - 1.00	-1.60	3.13

Note. The different sample sizes result from different excluded participants for each test. Here, the non-logarithmized reaction times are presented.

None of the tests were significantly correlated with the grade in German, $|r| \leq .14$, $p \geq .43$ and there was only one significant correlation between the validation tests and the grade in Mathematics, $r = -.33$, $p < .05$ (MRT). All other tests were less strongly correlated $|r| \leq .21$, $p \geq .29$. Experience with Rubik's cubes showed one marginally significant correlation with CC ($r = .27$, $p = .05$) but was not correlated with any other validation tests ($|r| \leq .16$, $p \geq .26$).

Table 16: Correlations between validation tests

	PFT	MRT	CC
MRT	.33*		
CC	.50***	.38*	
CT	-.42*	-.25	-.45**

Note. * $p < .05$; ** $p < .01$; *** $p < .001$. All correlations with CT based on $N = 44$ participants.

R-Cube-Vis Test for visualization: Descriptive results

The descriptive results of the single R-Cube-Vis levels showed the expected ordering (Table 17) with nearly the same difficulty of Level 4.1 and 3.1. Compared to Study 2, ACC-poss and the weighted accuracy (wACC) were higher for Level 4.2.c1 and similar to the results of Study 1, but with a lower accuracy than the chance level in Level 4.2.c2.

Table 17: Means (standard deviations) for the four R-Cube-Vis Test measures, ACC-all, ACC-poss, wACC and AUC, also for all levels separately

	ACC-all	ACC-poss	wACC	AUC
All levels	.83 (.08)	.83 (.09)	2.73 (0.52)	.93 (.06)
Level 4.1	.97 (.06)	.98 (.05)	3.56 (0.51)	.96 (.05)
Level 3.1	.98 (.04)	.98 (.06)	3.59 (0.49)	.94 (.04)
Level 4.2.p	.90 (.13)	.94 (.15)	3.16 (0.85)	.92 (.04)
Level 3.2.p	.80 (.16)	.88 (.20)	2.89 (0.86)	.91 (.04)
Level 4.2.c1	-	.76 (.25)	2.16 (1.15)	-
Level 4.2.c2	-	.42 (.22)	.96 (0.68)	-

R-Cube-Vis Test for visualization: Reliability and convergent and discriminant evidence of validity

The reliability estimates of the R-Cube-Vis measures were similar to Study 2, $\alpha = .76$ (accuracy over all items, ACC-all), $\alpha = .64$ (ACC-poss) and $\alpha = .90$ (wACC).

The correlations with the visualization test, PFT, and the MRT were significant for all measures and were descriptively larger for PFT than MRT, $.46 \leq r \leq .57$, $p < .01$ and $.34 \leq r \leq .45$, $p < .05$, respectively (Table 18). CC was also significantly correlated with three of four measures with effect sizes comparable to the correlations with the MRT. The correlations with CT were lower ($r \geq -.30$, $p \geq .05$) but were not significantly different from the correlations of all measures with the PFT $z \geq 1.30$, $p \geq .10$.

Table 18: Correlations of the R-Cube-Vis Test with visualization and spatial relations tests

	Measure	PFT	MRT	CC	CT
All levels	ACC-all	.46**	.41**	.31	-.16
	ACC-poss	.51***	.34*	.43**	-.28
	wACC	.57***	.42**	.45**	-.29
	AUC	.55***	.45**	.41**	-.30*

Note. * $p < .05$; ** $p < .01$; *** $p < .001$. All correlations with CT based on $N = 44$ participants. All correlations $|r| \geq .30$ are bold.

The grade in German was not significantly correlated with any measure, $|r| \leq .11$, $p \geq .44$. The correlations with the grade Mathematics were stronger, but only ACC-all achieved statistical significance, $r = -.36$, $p < .05$, all other correlations were weaker, $-.27 \leq r \leq -.18$, $p \geq .08$. The highest correlation with experience with Rubik's cube was found for wACC ($r = .26$, $p = .06$), all other measures were less correlated, $r \leq .16$, $p \geq .25$.

R-Cube-Vis Test for visualization: Testing conformity with the linear logistic test model (LLTM)

The 3-PL model with the fixed guessing parameter at .5 and the fixed discrimination parameter at 1 was more appropriate with moderate evidence than the RM according to AIC and BIC (RM: AIC = 947.11, BIC = 986.14; 3-PL: AIC = 942.94, BIC = 981.96). However, the differences in both criteria were small compared to the differences found in both previous studies, and therefore both models were interpreted as comparable to confirm the data structure. In the following analyses, the RM and the corresponding LLTM were considered to be consistent to Study 1 and 2.

Table 19: Results of Andersen's likelihood ratio test of RMs with three different splits

Split	χ^2	df	p	Infits between 0.5 and 1.5 [%]	PSR (RM)
ACC-poss	23.99	18	.16	100	.53
Between 1 and 2	23.10	19	.23	100	.71
Between 2 and 3	38.06	26	.06	100	.85
Between 3 and 4	16.31	23	.84	100	.92

There was no significant difference between β -estimates of the low and high ability groups according to Andersen's likelihood test for Rasch models (RMs), $p \geq .06$ (Table 19) for any of the variables (ACC-pos and three splits of wACC). All infits lay between 0.5 and 1.5 and all estimates of the person separation reliability (PSR) were acceptable, $r \geq .71$, except for the ACC-poss, $r = .53$ (Table 19). Plotting the β -estimates for both groups defined for the Andersen's tests against each other showed that all data points lie around the main diagonal (Figure 13).

Figure 13: Normalized β -parameters for RMs estimated from both subsamples used for Andersen's likelihood ratio test, below and above the mean.

Table 20: Comparison between RMs and corresponding LLTMs using log-likelihoods and reliability estimates for LLTMs for all splits

Split	χ^2	df	p	PSR (LLTM)
ACC-poss	16.55	22	.79	.52
Between 1 and 2	18.61	23	.72	.70
Between 2 and 3	23.36	25	.56	.85
Between 3 and 4	28.23	25	.30	.92

Comparisons between the RMs and the corresponding linear logistic test models (LLTMs) showed no significant differences based on the likelihood ratio tests for any split, $p \geq .30$ (Table 20). Both information criteria, AIC and BIC, showed strong evidence in favor of LLTMs compared to the corresponding RMs with $\Delta AIC > 10$ and $\Delta BIC > 50$. The reliability estimates (PSR) were nearly the same compared to the estimates of the RMs (Table 20). The visual inspection of the plots that assign the β -estimates of the RMs against β -estimates of the LLTMs revealed high consistency for corresponding values (Figure 14).

Figure 14: Comparison between normalized β -parameters for RMs and corresponding LLTMs.

Discussion

The aim of this study was the evaluation of an alternative procedure of the R-Cube-Vis short version, since both tested short version in Study 2 showed only moderate support for convergent and discriminant evidence of validity.

The correlations between the Paper Folding Test (PFT) and the Mental Rotation Test (MRT) as well as the correlations between the MRT and the Chronometric Test (CT) were lower than the respective correlations of the previous studies and below the lower bound of the expected intervals based on the current literature (e.g., Blajenkova et al., 2006; Kozhevnikov et al., 2002; Kozhevnikov & Hegarty, 2001; Miyake et al., 2001, see also Section 2). However, these unexpected correlations were interpreted as an artefact given the results of all three previous studies. All other correlations lay in their expected intervals. The negligible correlations with the validation tests and the grade in German confirmed the results of Study 1 and Study 2. The correlations with the grade in Mathematics were weaker than before, with the exception of the correlation with the MRT, but they were located in the expected interval too.

The correlation patterns found between all R-Cube-Vis measures and the validation tests were similar to Study 1, and stronger than those obtained in Study 2. They showed the expected pattern with strong correlations with the visualization test, PFT, weaker correlations with the MRT and the weakest correlation with the spatial relations test, CT. Only the correlations with the second spatial relations test, Comparison test (CC), were higher than expected and similar to the correlations with the MRT. However, the CC was also highly correlated with the PFT as the standard visualization test. Therefore, the correlation patterns of all measures support convergent and discriminant evidence of validity of the R-Cube-Vis Test as a visualization test.

The low to medium strength correlations with the external criteria - the grades in German and Mathematics - were comparable to the respective correlations found for the validation tests and are comparable to the results of the previous studies.

There is evidence for conformity with the linear logistic test models (LLTMs) as superior models compared to the corresponding Rasch models (RM). Hence, this supports the evidence of validity of the

internal structure. As in the previous studies, the small sample size limits the interpretation of Andersen's likelihood test and the likelihood ratio test because of the low power, but the alternative criteria support the model conformity.

The results of this study show that the changed procedure of the R-Cube-Vis short version improved the test with respect to convergent and discriminant evidence of validity and is more suitable than the procedure applied in Study 2.

Study 4: Disentangling type- and item-specific characteristics of gaze patterns

When looking at parts of a visual item during the attempt to solve the item, the gaze patterns are thought to be determined by the cognitive processes that are necessary to solve the item. These type-specific characteristics are assumed to result from “top down” processing. However, there are visual features of the item that are independent of the demanded cognitive processes and which are therefore task-irrelevant. But they may affect the gaze pattern as well by “bottom up” processing resulting in item-specific characteristics (see Section 2). In addition, there might be random eye movements, which are considered as unsystematic error and will be ignored in the following analysis. In order to identify cognitive processes that are strategically applied to solve an item (“top down”), the gaze patterns that indicate processing of the crucial visual information have to be disentangled from gaze patterns driven by visual features in “bottom-up” fashion that “capture” the eyes but that are task-irrelevant. Hence, the goal of this study is to disentangle the type-specific characteristics of the recorded gaze-patterns from the item-specific characteristics.

The R-Cube-Vis Test provides an approach to resolve this issue. Its six levels of difficulty are considered to be *homogeneous* item sets. All items of a given level are constructed based on the same transformation principle (called the *type* of a certain level). For instance, the items of Level 4.1 (Figure 17, p. 70) are constructed such that there is a one-step rotation applied to one axis of the cube on the right. It is assumed that processing all items of this level will require corresponding essential mental transformation steps to solve them successfully. For the items of Level 4.1, identifying the rotated axis and mentally performing a back rotation may be necessary steps. However, the particular axis, or the side of the cube where the rotation is visible, as well as the rotation direction, vary between the items of Level 4.1. These differences in visual appearance are assumed to be independent of the required cognitive processing steps. The basic idea is that gaze patterns that are equal over all items of a specific level reflect the cognitive processes (top-down) that are applied to solve an item and are interpreted to be *type-specific*. In contrast, gaze patterns that vary between items of the same level reflect visual features of an item that are not relevant for solving the task (“bottom-up”) and are interpreted to be *item-specific*. Therefore, these two kinds of gaze patterns (type-specific and item-specific) can be distinguished by fixation-based measures of all items of the R-Cube-Vis Test that belong to the same difficulty level, where all items have the same type.

In accordance with existing eye tracking research, it was expected “bottom-up” processes should be more observable on fine-scaled areas of interest (AOIs), i.e., single elements of the cubes, because of visual attributes of these features that might lead the gaze (see Ivie & Embretson, 2010). In contrast, large-scaled AOIs based on summarizing many fine-scaled AOIs, e.g., the whole cube, might result in effects that are independent of these specific attribute effects (see Snow, 1980).

The fine-scaled AOIs of the single elements of the cubes are coded with respect to the construction principle of the item type. For example, the right cube of each item of Level 4.1 shows one side with no rotated element (Figure 17, p. 70). However, for each cube of Level 4.1 this side can be the left side, the top side, or the right side of the right cube. For all items, however, this side is labeled equally, because it is the side with no rotated element. This way to code the AOIs is termed “function coding” here. When fixations on functionally coded AOIs *differ* between homogenous items of the *same type*, there must be item-specific visual features that lead the eyes in different ways for each item. Furthermore, particular visual features may “capture” the eyes in predictable ways across items of a type independent of their function for solving the item. This would be the case if, for instance, a particular color pattern or a particular actual side of a

cube would always be fixated for some time. Understanding these purely visual item features across items of a type requires another coding of the AOIs with regard to the actual visual places and characteristics rather than with regard to their function for the transformations. This alternative coding of the AOIs considers the actual place (left, top, or right side) and is labelled “location coding” here.

The specific fixation-based indicators, which were considered in the following analyses, focus either on small parts of a side of a cube that are functionally relevant for mental transformations (fine-scaled AOIs), but also comprise larger AOIs such as one side of a cube or a complete cube. In addition, the indicators may encompass patterns of subsequent visits on AOIs. In the present study, four indicator classes were considered: (1) Fixations on pre-defined, functionally coded parts on a single cube (fine-scaled), with the functional relevance based on the construction principle of the item type, (2) fixations on the complete cube, i.e., summing up over all fine-scaled AOIs on a single cube (large-scaled), (3) sequences of fixation switches between the left and the right cube, i.e., summing up over time (large-scaled), and (4) specific viewing patterns on the fine-scaled, functionally coded AOIs that were assumed to be necessary to solve an item according to its type (fine-scaled). Item-specific effects should be found for fixations-based measures of fine-scaled AOIs (Indicator-Class 1 and 4), whereas type-specific effects were expected for measures of large-scaled AOIs (Indicator-Class 2 and 3).

In addition, an attempt was made to understand visual item features that may determine item-specific differences in fine-scaled, functionally coded AOIs between items of the same type, i.e., to understand how visual features drive “bottom-up” effects in gaze patterns that are unrelated to solving the item. To this end, the location coded AOIs were interpreted with respect to the single items. A better understanding of the “bottom up” effects is important. If the strategic “top down” transformation processes have to be identified and described with gaze patterns, then the “bottom up” gaze patterns have to be controlled for.

The second part of Study 4 was the explorative analysis of all fixation-based measures with respect to their potential to indicate the correctness of the answer of single items utilizing logistic regression models. These results would provide insights into the cognitive processes indicated by the fixations-based measures that might be necessary, or at least beneficial, to solving items of a certain level of difficulty.

Method

Participants

All $N = 69$ participants were students at the University of Mannheim, were not color-blind, and received course credit. Seventeen participants had to be excluded based on the same criteria as in the previous studies and due to erroneous eye tracking data. The final sample consisted of $N = 52$ participants (42 female, 10 male) with mean age $M = 20.90$ years ($SD = 3.87$ years) ranging from 18 years to 39 years.

Materials

The short version of the R-Cube-Vis Test was conducted following the same procedure as in Study 3. Before each item, a cross was presented in the middle of the screen for 1 second, on which the participants had to fixate. The cubes of each following item were placed to the right and left of this fixation cross. Furthermore, participants had to provide information about sex, age, and study major. As in Study 3, the test was computer-based with self- and system-administered parts.

Apparatus

The R-Cube-Vis Test was presented on a 23” screen (aspect ratio: 16:9, resolution: 1920 x 1080 pixels) embedded in an eye tracking unit with the Tobii TX300 remote eye tracker with 300Hz recording frequency. The eye tracker was connected to the presentation software E-Prime 2.0 (Psychology Software Tools Inc., 2012) with

the “Extensions for Tobii” (Psychology Software Tools Inc., 2011). Fixations were computed by the adaptive event detection algorithm of Nyström and Holmqvist (2010). The algorithm was applied with an upper bound at 200ms/deg to limit the adaptive event detection thresholds if the data are too noisy (Fehringer, 2018). The algorithm was implemented in python (van Rossum, 1995) with the packages `pandas` (McKinney, 2011), `numpy` (van der Walt, Colbert, & Varoquaux, 2011), and `scipy` (Jones, Oliphant, Peterson, & others, 2001). The recorded pupillometric data were not considered in the following analyses. However, they were used to optimize a pupillary-based indicator for cognitive workload, the Index of Pupillary Activity (Duchowski et al., 2018), with respect to the difficulty levels of the R-Cube-Vis Test (Fehringer, 2020).

Procedure

After the instruction, the eye tracker was calibrated for each participant. They then had to perform the R-Cube-Vis Test. At the end, participants had to answer the questionnaire. Before and after the R-Cube-Vis Test, a fixation task with nine fixation crosses was performed to provide evidence for comparable accuracies of the eye tracker at the beginning and the end of the experiment. All tasks were presented with the presentation software E-Prime 2.0 (Psychology Software Tools Inc., 2012).

Analyses

Areas of interest (AOIs) were defined to cluster fixations on the same regions. These AOIs were defined either fine-scaled for regions with the same color (Figure 15), i.e., one AOI for each feature on a cube (Indicator-Class 1), and by combining fine-scaled AOIs to more meaningful AOIs as whole cubes (Indicator-Class 2). As described above, the AOIs were function coded and, if possible, also location coded.

Figure 15: Cubes with fine-scaled AOIs outlined in light green.

For each of these defined AOIs, separately for each item, the relative number of fixations was considered. The relative number of fixations of an AOI X was defined as the number of fixations on X divided by the number of all fixations on one of the two cubes.

Second, entropy values (Shannon, 1948) for each level were computed as an indicator of how chaotic or systematic the fixation patterns were (Indicator-Class 3). Entropy, $H(X)$ is computed by

$$H(X) = - \sum_{i=1}^s \pi_i \sum_{j=1}^s p_{ij} \log(p_{ij})$$

with the transition matrix X that contains in each cell (i, j) the probability that a fixation on AOI i is followed by a fixation on AOI j . The entry in each cell (i, j) of X is labeled by p_{ij} , s is the number of AOIs and π_i is the probability of occurrence of AOI i . The log-function is the logarithm to the base e (natural logarithm). Simplified, the lower the entropy value is, the more systematic and less chaotic is the AOI sequence. In order to take different fixation durations into account, fixations were split up into 200ms chunks, i.e., an origin fixation with duration of 560ms was split up into three fixations (200ms + 200ms + 160ms). That means a

fixation sequence ABAAB with fixation durations 560ms, 180ms, 320ms, 100ms, 210ms would result in a sequence of AAABAAABB with theoretical durations (200ms + 200ms + 160ms), 180ms, (200ms + 120ms), 100ms, (200ms + 10ms). Due to the short fixation sequences, especially at the easiest levels, entropy values were analyzed for fixations sequences only differentiating between left and right cube, because some of the fine-scaled AOIs were not fixated at all and would lead to many empty entries in the transition matrix X . Hence, the entropy values contain only fixations on the left and the right cube and base, therefore, on indicators of Indicator-Class 2.

Third, fixation sequences per item were analyzed in order to find indicators of cognitive processes that are potentially necessary to solve the items (Indicator-Class 4), i.e., back rotation of the rotated element (Figure 16a) and comparison of corresponding parts of the right cube (Figure 16b, c). Back rotation means that participants tried to imitate the required rotation of the rotated element to solve the cube. Comparison of corresponding parts means that participants compared one side of a rotated element with corresponding non-rotated elements of another cube side. To this end, the fixation sequences per item were searched for bi-grams, a special case of n -grams with $n = 2$. A n -gram is a sequence of n parts, which is usually searched for in a bigger, ordered sequence, e.g., the sequence of n letters is searched for in a given text. In the current study, two consecutive fixations (bi-grams), which might indicate the described cognitive processes, were searched for in the fixation sequences of each item and each participant. For each rotated element, three bi-grams were defined. (1) Back rotation was defined by the first fixation on the “old” color of the rotated element (an “old” color is a color that appears on both cube, left and right, i.e., this color would be still visible if the right cube is solved) and a second fixation on the “new” color (this color is not visible on the left cube, this color only appears on the right cube due to the rotation) of the rotated element (Figure 16a, first fixation should be on the red part and the second fixation on the green part). For the comparison of corresponding parts, a differentiation between the old and the new color of the rotated element was considered. One fixation should be (2) either on the old color (Comparison I) or (3) on the new color (Comparison II) of the rotated element. The second fixation should always be on the not-rotated elements of the other cube side (Comparison I: Figure 16b, fixation on the blue part of the rotated element is followed by a fixation on the blue part of the not-rotated element or vice versa. Comparison II: Figure 16c, fixation on the brown part of the rotated element is followed by a fixation on the green part of the not-rotated element or vice versa). One should note that for the bi-grams that indicate rotation, the fixation sequence is ordered, whereas for the bi-grams indicating comparisons, the ordering of the two fixations is arbitrary. For each bi-gram, each item, and each participant, a variable indicates whether the respective bi-gram occurs (0: no bi-gram; 1: at least one bi-gram).

Figure 16: Potential gaze patterns that are necessary to solve the items: a) back rotation, b) comparison between the rotated part with the old color and the corresponding non-rotated part (Comparison I) and c) comparison between the rotated part with the new color and the corresponding non-rotated part (Comparison II)

The entropy values as well as the bi-grams were computed using `python` (van Rossum, 1995) with the packages `pandas` (McKinney, 2011), `numpy` (van der Walt et al., 2011), and `scipy` (Jones et al., 2001). All plots were created by the R-package `ggplot2` (Wickham, 2009).

First, the indicators of the four indicator classes were examined with respect to type-specific and item-specific properties. To this end, the means and the 95%-confidence intervals of the relative number of

fixations of each indicator of Indicator-Class 1 and 2 as well as the values of the entropy (Indicator-Class 3) and the bi-grams (Indicator-Class 4) were compared between items of the same level of difficulty. Item-specific properties exist if the five possible items of a specific level differ in the considered means. This holds for only the function and not the location coded fine-scaled AOIs. If all means are equal for a specific indicator and a single level of difficulty, this indicator was interpreted as type-specific. These analyses were only performed descriptively and not with significance statistics due to the new approach in these analyses and the absence of previous knowledge about the gaze patterns of the utilized stimuli and about the differentiation between type- and item-specific characteristics.

Item-specific characteristics of the fine-scaled AOIs, indicated by differences between the items of the same level, were inspected regarding specific features such as the position of the rotated elements or the axis around which the elements were rotated utilizing the location coded AOIs. Based on the identified item configuration and the corresponding means of the considered indicator, assumptions were derived to explain the different means. For example, if the rotated element was rotated around the horizontal axis, it could be that the participants look more often at the rotated element than if the rotated element was rotated around the vertical axis.

Third, all indicators were examined regarding their ability to differentiate between correctly and incorrectly solved items (item accuracy). Therefore, logistic regression models were analyzed as prediction models with the single tested indicators as independent variable and item accuracy as dependent variable. Item accuracy can either be 0 (correctly solved item) or 1 (incorrectly solved item). All indicators were z-standardized to control for item-dependent properties. The significance level was set to $\alpha = .05$ for all tested models. However, these models were only computed for the two most difficult levels (Level 4.2.c1 and 4.2.c2), because all other levels had too few incorrectly solved items. For both medium levels (Level 4.2.p and 3.2.p) differences between the means of the correctly and incorrectly solved items were descriptively analyzed. Such descriptive comparisons could not be performed for the easiest levels (Level 4.1 and 3.1), because some items were solved incorrectly only by one or even zero participants.

Similar as before, correctly solved items should be related with lower entropy values (indicating higher systematic viewing patterns) and with a higher occurrence of bi-grams assuming that these fixation sequences reflect cognitive processes that are necessary for solving the items.

Results

The expected differences between the levels could be found for ACC-poss (accuracy over all possible items) and with regard to reaction times, number of fixations, and fixation duration (Table 21). Generally, the more difficult the levels were, the longer were the reaction times and the fixation durations, and the larger was the number of fixations. However, the results were comparable between Level 4.1 and 3.1 and between Level 4.2.p and 3.2.p but differed between Level 4.2.c1 and 4.2.c2.

Table 21: Mean (standard deviation) of reaction times, number of fixations and fixation duration separated for each level

Level	N	ACC-poss		Reaction time		Number of fixations		Fixation duration	
				M (SD)		M (SD)		M (SD)	
Level 4.1	49	.95	(.17)	5503.55	(3889.37)	11.79	(7.81)	3877.15	(3214.77)
Level 3.1	47	.93	(.15)	5540.35	(4380.15)	12.28	(9.58)	3879.25	(3559.94)
Level 4.2.p	52	.86	(.21)	10085.76	(7198.14)	20.60	(15.51)	7700.26	(5969.93)
Level 3.2.p	52	.81	(.22)	10262.76	(5970.52)	21.69	(13.97)	7819.22	(4804.38)
Level 4.2.c1	52	.72	(.26)	11424.63	(7543.50)	24.23	(18.97)	8810.05	(6562.00)
Level 4.2.c2	52	.40	(.29)	14089.39	(9471.10)	29.46	(21.15)	10492.14	(7485.59)

Differentiation between item-dependent and item-independent properties

Single cube features and whole cubes

Only a few selected results are presented due to space restrictions. For demonstration purposes, the focus of the following analyses lies on the left cube (Indicator-Class 2) and its three sides (Indicator-Class 1). An investigation was made whether the indicators of the three sides of the left cube change depending on the corresponding side of the right cube. For example, an analysis was performed whether the number of fixations on the side of the left cube that corresponds to the side of the right cube showing no rotated elements differs from the number of fixations on a side of the left cube that corresponds to a side of the right which shows the rotated element. In the following only the results of the easiest level, Level 4.1, the medium difficulty level, Level 3.2.p, and the most difficult level with accuracy larger than chance level, Level 4.2.c1 are presented.

Figure 17: Possible items of the R-Cube-Vis (short) test of Level 4.1

Figure 18: Example of the labeling of item 4.1_8

The considered items of Level 4.1 are displayed in Figure 17. The AOI that covers the complete left cube is labeled with “L”. All single sides of the left are labeled according to the corresponding side of the right cube (function coded, see the example coding for item 4.1_8 in Figure 18). If the side of the right cube shows no rotated element, it is labeled with “L.not”, e.g., the brown side of item 4.1_8 (Figure 18) and the red side of item 4.1_9 (Figure 17). If the side of the right cube shows the side of the rotated element with the color

that cannot be seen on the left cube (“new color”), the corresponding side of the left cube is labeled with “L.new”, e.g., the red side of item 4.1_8 (Figure 18) and the blue side of item 4.1_16 (Figure 17). If the side of the right cube shows the side of the rotated element with the color that is also be shown on the left cube (“old color”), the corresponding side of the left cube is labeled with “L.old”, e.g. the yellow side of item 4.1_8 (Figure 18) and the blue side of item 4.1_19 (Figure 17). The location coded AOIs were labeled with “L.left” for the left side, “L.top” for the top side, and “L.right” for the right side.

As expected, the descriptive analysis of the fine-scaled, function coded AOIs (Indicator-Class 1) revealed item-specific effects with a different relative number of fixations for the five items. The large-scaled AOI (Indicator-Class 2), the left cube, showed a similar relative number of fixations for all items, except for item 4.1_19 (Figure 19) and was therefore interpreted as type-specific. However, the differences between the fine-scaled AOIs disappeared nearly completely if the AOIs were location coded (Figure 19). For all items, the right side of the left cube that is closest to the middle of the item (where the fixation cross was presented directly before the item was presented) had the highest relative number of fixations of all sides of the left cube followed by the top side, which is the second closest side to the middle. The left side of the left cube with the largest distance to the middle showed the smallest relative number of fixations. The reason might be that participants have to fixate always in the middle of the screen before the item appears and that the cubes of each item are placed to the left and to the right of the screen center. Participants might have fixated more frequently on cube parts the closer they were to the center. However, item 4.1_9 was an exception also in this analysis with a lower relative number of fixations on the top side than was the case for the other items. One reason might be that this item is the only one for which the two top layers are not rotated. Therefore, participants might have fixated on the top layer to a lesser extent, because the attention is only on the lower parts of the cube.

Figure 19: Mean relative number of fixations of the AOIs of the left cube as well as its single sides, function and location coded in Level 4.1. The sum of the three sides is equal to the value of the whole cube (AOI L). The error bars indicate the 95%-confidence intervals.

The items of Level 3.2.p are presented in Figure 20. As for Level 4.1, “L.not” refers to the side of the left cube corresponding to the side of the right with no rotated elements. The other two sides with rotated elements cannot be semantically discriminated. On both sides one rotated element shows a new color and the other rotated element shows an old color. Therefore, the numbers for these sides were randomly chosen (Figure 20). In the analyses, both sides were considered together and labeled with “L.rot”. The location coded AOIs L.left, L.top, and L.right are defined as above.

Figure 20: Possible items of the R-Cube-Vis (short) test of Level 3.2.p

The descriptive analysis of the large-scaled AOI (Indicator-Class 2), the left cube, showed a type-specific pattern of the relative numbers of fixations with nearly the same values for all items in Level 3.2.p (Figure 21). In contrast to Level 4.1, the fine-scaled function coded AOIs (Indicator-Class 1) were also very similar for all items, whereas the location coded AOIs differed across the items (Figure 21). However, the location coded AOIs showed only one side with higher relative number of fixations for each item compared to all other sides (Figure 21). Each of these specific sides is one of the two sides that correspond to one of the sides of the right cube showing both rotated elements (Figure 20). This pattern seems to follow a simple rule: If the right side of the left cube (nearest to the middle of the screen) corresponds to a side of the right cube that shows the two rotated elements, then this side has the highest relative number of fixations. If this is not the case, then the both other sides of the left cube correspond to the both sides of the right cube showing the rotated elements and then the second closest side to the middle, the top side, has the highest relative number of fixations. However, item 3.2_1 is an exception with the highest relative number of fixations on the left side of the left cube, which may have been caused by the red color of one of the rotated elements leading the gaze. These results suggest that for these items, both processes, top-down and the bottom-up, are important with respect to gaze patterns and that the resulting gaze patterns have type- and item-specific characteristics.

Figure 21: Mean relative number of fixations of the AOIs of the left cube (AOI L) as well as the sides corresponding to the non-rotated side of the right cube (L.not) and both other sides combined (L.rot) function coded and all three sides location coded in Level 3.2.p. The average relative number of fixations per side is presented at the AOI L.rot (average). The sum of the three sides is equal to the value of the whole cube (AOI L). The error bars indicate the 95%-confidence intervals.

The considered items of Level 4.2.c1 are shown in Figure 22. Here, the labels of the left cube differ from before. “L.cross” refers to the side of the left cube corresponding to the side of the right cube on which both elements cross, e.g., the green side of item 4.3_10 (Figure 22). “L.point” is the label of the left cube’s side corresponding to the side of the right cube showing one rotated element with a “dot”, e.g., the white side of item 4.3_23 (Figure 22). “L.line” refers to the side of the left cube that corresponds to the side of the right showing one rotated element with no dot, e.g., the yellow side of item 4.3_10 (Figure 22). Also in this case, the location coded AOIs L.left, L.top, and L.right refer to the left side, the top side, and the right side, respectively.

The expected type-specific pattern of the relative number of fixations for the large-scaled AOI, the left cube, were also revealed by the descriptive analysis of Level 4.2.c1. The relative number of fixations of the fine-scaled AOIs differs between the items with function as well as with location coding (Figure 23). However, some systematics might be recognized. In four of the five items, the side of the left cube that corresponds to the side of the right cube showing the cross had the highest relative number of fixations (the exception is item 4.3_17). One reason could be that this side (AOI L.cross) corresponds to the only side of the right cube that shows both rotated elements. Hence, participant might have focused more on this side to understand how both elements are rotated. Also, the relative number of fixations on the left side of the left cube, which has the largest distance to the middle, was always the lowest, except for item 4.3_17, that shows the cross on this side. Similar as in Level 3.2.p, the gaze pattern seems to be led by top-down as well as by bottom-up processes and, therefore, contain type- and item-specific characteristics.

Figure 22: Possible items of the R-Cube-Vis (short) test of Level 4.2.c1

Figure 23: Mean relative number of fixations of the AOIs of the left cube as well as its single sides, function and location coded in Level 4.2.c1. The sum of the three sides is equal to the value of the whole cube (AOI L). The error bars indicate the 95%-confidence intervals.

Entropy

In contrast to the relative number of fixations on single cube features or even on the complete cube, the descriptive analysis showed that the entropy values (Indicator-Class 3, summing up over time) seemed to be more stable for items within the same difficulty level (Figure 24). The value ranges within a single item level were close and went from .03 (Level 4.2.c1) to .08 (Level 4.2.c2) for all levels indicating type-specific behavior. This result confirms the assumption that more broadly defined fixation-based indicators (large-scaled) are more stable for different items from the same difficulty level and are, therefore, type-specific.

The between-level differences showed higher values for easier levels and lower values for the most difficult levels. Level 3.2.p seemed to be an exception with higher entropy values than Level 4.2.p and with entropy values comparable to Level 4.1 and 3.1. One reason might be that for items of Level 3.2.p in particular it was not obvious which elements had to be back rotated by observing only the right cube, because all elements have different colors on a single side (Figure 20). Therefore, it was comparatively more helpful to look at the left cube than it was for each other level. However, further studies with more detailed analyses of the fixation sequences are needed to confirm this finding.

Figure 24: Entropy values for switches between the AOI L, the left cube and AOI R, the right cube. Each point indicates a single item of the respective difficulty level.

Bi-grams

The relative frequencies of the analyzed bi-grams differed between different levels as well as between single items within the same level (see the plots for Level 4.1, Figure 25, Level 3.2.p, Figure 26, and Level 4.2.c1, Figure 27). This result confirms that the bi-grams (Indicator-Class 4) can be interpreted to be item-specific. However, there were some exceptions showing a more type-specific pattern, such as Comparison II of Level 4.1 (Figure 25), and all bi-grams of Element 1 of Level 4.2.c1 (Figure 27). Some of the results might be explained by visual item features of the items. For example, the items of Level 3.2.p might be grouped into two groups regarding the occurrence of the Rotation (Figure 26). Items 3.2_1, 3.2_10, and 3.2_18 had a higher occurrence than the items 3.2_4 and 3.2_20. A reason might be that both rotated elements of the first three items (items 3.2_1, 3.2_10, and 3.2_18) are next to each other, whereas both rotated elements of the last two items (3.2_4 and 3.2_20) have the non-rotated element placed between them (Figure 20). In both cubes, the non-rotated element is in the middle. The participants, therefore, had to switch their gaze fixation from the right front to the left back.

Figure 25: Mean occurrence rate of the three bi-grams in Level 4.1 separated for each item. Comparison I refers to the comparison between the old color of the rotated element and the corresponding parts, Comparison II refers to the comparison between the new color and the corresponding parts, whereas Rotation means the gaze switch from the old to the new color of the rotated element (for details see Section “Analyses”). The error bars indicate the 95%-confidence intervals.

Figure 26: Mean occurrence rate of the three bi-grams in Level 3.2.p separated for each item. Comparison I, Comparison II, and Rotation are defined as above combined for both elements. The error bars indicate the 95%-confidence intervals.

Differentiation between correctly and incorrectly solved items

Differences between correctly and incorrectly solved items were analyzed descriptively regarding the number of fixations on all defined AOIs (Indicator-Class 1 and 2), the entropy (Indicator-Class 3), and the bi-grams (Indicator-Class 4). However, this differentiation was not considered for Level 4.1 and 3.1, because there were too few incorrectly solved items in these levels (Level 4.1: $1 \leq N \leq 5$; Level 3.1.p: $2 \leq N \leq 7$).

Figure 27: Mean occurrence rate of the three bi-grams in Level 4.2.c1 separated for each item. Comparison I, Comparison II, and Rotation are defined as above. Element 1 refers to the first element that has to be rotated to receive the right cube from the left cube. Element 2 refers to the other element. The error bars indicate the 95%-confidence intervals.

The number of incorrectly solved items per level was still low for the two levels with medium difficulty (Level 4.2.p: $4 \leq N \leq 10$; Level 3.2.p: $6 \leq N \leq 12$), therefore, no statistical models were computed and the data were only analyzed descriptively. For both levels, Level 4.2.p and Level 3.2.p, there seemed to exist a type-specific pattern of the AOI L.rot that combined the two sides of the left cube that correspond to the two sides of the right cube with rotated elements. The relative number of fixations was always lower for incorrectly than for correctly solved items with differences $.05 \leq M_{\text{correct}} - M_{\text{incorrect}} \leq .08$ (Level 4.2.p) and $.02 \leq M_{\text{correct}} - M_{\text{incorrect}} \leq .08$ (Level 3.2.p). This could mean that the comparison of the rotated elements of the right cube with the corresponding sides of the left cube is helpful to solve these items and would be a type-specific property.

The logistic regression models showed an effect for the left cube (AOI L) in Level 4.2.c2, $\chi^2(1) = 5.45$, $p < .05$; $b = -0.31$, $p < .05$, OR (odd ratio) = 0.73, [0.56, 0.95] that, however, seems to be traced back to the AOI L.point, $\chi^2(1) = 7.61$, $p < .01$; $b = -0.40$, $p < .01$, OR = 0.66, [0.51, 0.85]. In contrast to Level 3.2.p, a higher relative number of fixations was indicative for incorrectly solved items. A similar effect for the left cube (AOI L) could be found in the descriptive analysis for Level 4.2.c1. For four out of five items of Level 4.2.c1, the values were higher for incorrectly solved items, $.03 \leq M_{\text{incorrect}} - M_{\text{correct}} \leq .11$, such as for Level 4.2.c2.

A reason for the reversed pattern of Level 4.2.c1 and 4.2.c2 compared to Level 4.2.p and 3.2.p might be certain characteristics of the items. For all levels, the rotated elements have to be compared with the non-rotated elements in order to decide in which direction they have to be rotated back. For all levels, the information of the non-rotated elements can be gained from the left cube, but also from the non-rotated elements of the right cube. For Level 4.2.p and particularly for Level 3.2.p recognizing the rotated elements is not obvious by only observing the right cube, because of the two parallel rotated elements visual on two sides of the right cube (Figure 20). Hence, it might be that detecting the non-rotated parts on the right cube is more error-prone than to detect these parts on the left cube. In contrast, for Level 4.2.c1 and 4.2.c2, it is easier to recognize the non-rotated elements of the right cube, because only on one side of the cube there are two rotated elements (on the other two sides, there is only one rotated element) and these elements are crossed. Therefore, they are obviously recognizable as rotated elements and can easily be distinguished from the non-rotated elements (Figure 22). Hence, it might be that if participants focused more on the right cube than on the left one, they could solve the items correctly more easily because they could reduce the error-prone switching between the left and the right cube.

The descriptive analyses of the entropy values of Level 4.2.p and 3.2.p showed in eight of ten items lower values for the incorrectly solved items (Level 4.2.p: $.05 \leq M_{\text{correct}} - M_{\text{incorrect}} \leq .09$; Level 3.2.p: $.05 \leq M_{\text{correct}} - M_{\text{incorrect}} \leq .15$). This result suggests that if participants had a more chaotic viewing pattern, they had a higher chance to solve the items correctly. However, due to the small samples of the incorrectly solved items, this pattern has to be interpreted with caution. Only the linear logistic regression model of Level 4.2.c2 was significant, $\chi^2(1) = 4.39$, $p < .05$; $b = -0.30$, $p < .05$, $OR = 0.74$, $[0.55, 0.98]$ indicating that lower entropy (more systematic viewing patterns) predicts a higher probability to solve the item correctly. This pattern was reversed for Level 4.2.p and 3.2.p. The differences in the mean entropy values of the single items were $0.04 \leq M_{\text{incorrect}} - M_{\text{correct}} \leq .09$. The order of the entropy values of the incorrectly and correctly solved items for Level 4.2.c1 was ambiguous.

The different results of Level 4.2.p and 3.2.p compared to Level 4.2.c2 might be explained in the same way as before. For Level 4.2.p and 3.2.p, it might have been more promising to solve the items correctly by switching – maybe in a chaotic way – between the left and the right cube. For the most difficult level, Level 4.2.c2, a more systematic switching between the left and the right cube might have been more successful.

The descriptive analysis of the bi-grams of the medium levels (Level 4.2.p and 3.2.p) showed no systematic difference between correctly and incorrectly solved items. In contrast, for some of the bi-grams, significant results could be found for the most difficult levels: for Level 4.2.c1, Rotation of the second element, $\chi^2(1) = 6.75$, $p < .01$; $b = 0.40$, $p < .05$, OR (odds ratio) = 1.49, 95%-confidence interval = $[1.10, 2.06]$; Comparison II of the first element, $\chi^2(1) = 4.91$, $p < .05$; $b = 0.33$, $p < .05$, $OR = 1.40$, $[1.04, 1.93]$; Comparison II of the second element, $\chi^2(1) = 11.27$, $p < .001$; $b = 0.47$, $p < .001$, $OR = 1.60$, $[1.22, 2.12]$ and for Level 4.2.c2, Comparison I of the second element, $\chi^2(1) = 10.87$, $p < .001$; $b = 0.43$, $p < .01$, $OR = 1.53$, $[1.19, 1.99]$. All significant models suggest that the occurrence of the respective bi-grams raises the probability of solving the item correctly. The descriptive analysis considering all bi-grams suggested a general effect that the bi-grams can differentiate between correctly and incorrectly solved items for Level 4.2.c1 and 4.2.c2. These results seem to be independent of the specific value-level of the relative frequency of the single item. They suggest that if the cognitive processes indicated by the bi-grams are performed by the participant, the respective item is solved correctly with a higher chance. Although the absolute values differed between items and can be interpreted as item-specific, the differences between correctly and incorrectly solved items suggest a type-specific pattern. However, for some items, the absolute values are generally low for incorrectly as well as for correctly solved items.

General discussion

Eye tracking based measures can deliver additional information about how participants solve a certain test item that goes beyond the information gained by accuracy and reaction times. Therefore, the goal of the present studies was to analyze the potential of different eye tracking measures to indicate specific cognitive

processing steps, which might be able to predict item answers. However, in a previous step type- and item-specific characteristics had to be disentangled. The studies were conducted in the well-established domain of spatial thinking. A new test, the R-Cube-Vis Test, was developed and validated (Study 1 to 3). The test measures the main factor of spatial thinking, visualization. It overcomes the drawbacks of established visualization tests with respect to the usage of eye tracking. Study 4 demonstrated the application of these stimulus materials with various fixation-based measures.

Development and validation of the R-Cube-Vis Test

Existing tests of visualization, such as the Paper Folding Test (PFT, Ekstrom et al., 1976a) or the Mental Rotation Test (MRT, Vandenberg & Kuse, 1978), comprise a small number of heterogeneous and complex items. Relevant features overlap, and the tests are administered as paper-pencil-versions. These properties hamper the ease of use of eye tracking. Therefore, a new visualization test, the R-Cube-Vis Test, was developed to overcome these drawbacks to be suitable for eye-tracking techniques. The items of the R-Cube-Vis Test were rationally constructed in accordance with the definition of visualization (Carroll, 1993). The R-Cube-Vis Test consists of six difficulty levels with 24 items (twelve possible and twelve impossible) per level in the long version and ten items (five possible and five impossible) per level in the short version. Four accuracy-based measures were considered in both versions, accuracy over all items (ACC-all), over all possible items (ACC-poss), weighted accuracy (wACC), and the Area Under the Curve (AUC). All measures delivered comparable results in the validation studies, but ACC-poss and wACC were the only measures that could differentiate between the two most difficult levels.

The reliability estimates based on Cronbach's alpha for the long version (Study 1) of all measures were very good. Lower reliability estimates were found in the short versions (Study 2 and 3) but were assumed to be an artefact caused by the binary format (Brogden, 1946a, 1946b; Sun et al., 2007) combined with the small number of items.

The validity of the R-Cube-Vis Test was considered according to the standards formulated by American Educational Research Association et al. (2014). The content-oriented evidence of validity of the R-Cube-Vis Test is given by its construction according to the definition of visualization. There is also evidence of validity regarding cognitive processes assuming that the participants mentally rotated the single elements to solve the items. This assumption is additionally supported by the results of the analyzed bi-grams (fixation sequences based on two fixations) in Study 4. Here, three kinds of bi-grams were considered that might indicate the back-rotation of the rotated elements as well as comparisons between one side of a single rotated element and the corresponding parts of the not-rotated elements. The cognitive processes presumably indicated by these bi-grams were assumed to be necessary to solve the items in the intended manner with the so-called visualization strategy. For the two most difficult levels it could be shown that such cognitive processes indicated by those bi-grams occur more frequently for the correctly solved items than for the incorrectly solved items. This supported the evidence of validity regarding cognitive processes.

Convergent and discriminant evidence of validity could be shown in Study 1 and 3 for all measures regarding the correlations with the PFT for visualization, the chronometric test (CT, Jansen-Osmann & Heil, 2007), the Würfelaufgaben (form A) from the Intelligenz-Struktur-Test (IST-WA, Amthauer et al., 1999), and the Cube Comparison test (CC, Ekstrom et al., 1976a) for spatial relations as well as the MRT in the middle of the continuum (Pellegrino et al., 1984). In both studies, all correlations with the PFT were strong and descriptively larger than the correlation between the PFT and MRT. The correlations with the CT were weak to medium. The correlations with the MRT and the second spatial relations test (IST-WA in Study 1 and CC in Study 3) lay in between and were comparable with each other. Whereas the correlations with the MRT were expected because the MRT is located in the middle of the continuum between visualization and spatial relations, the correlations with the second spatial relations test were descriptively larger than expected. The original assumption was that the correlations would be the same as with the other spatial relations test, CT. The reason for the stronger correlations with the second spatial relations test might be that these tests (IST-WA and CC) are closer to the visualization end of the continuum than the CT. This idea is supported by the descriptively larger correlations of these tests with the PFT and MRT in Study 1 and

3. However, the correlation patterns of all measures confirm the expected patterns regarding convergent and discriminant evidence of validity. The weaker convergent and discriminant evidence of validity in Study 2 can be explained by the testing procedure. Changing the testing procedure in Study 3 revealed the expected correlation pattern. The differences between the four accuracy-based measures are small and only descriptive. A potential effect must be tested in a follow-up study.

One important requirement for the R-Cube-Vis Test with respect to the ease of use of eye tracking was the creation of many homogenous items that should be equal regarding their difficulty on the measured ability dimension. Therefore, all items from one homogenous group should be on the same difficulty level. To examine the evidence of validity of the internal structure, the R-Cube-Vis stimulus materials were tested for conformity with the linear logistic test model (LLTM) the dichotomous measure ACC-poss and for the three splits of wACC. The splitting was necessary because of the bi-modal distribution of the single values of each item. The LLTM parameter indicated characteristics of the cubes that were potentially relevant for solving the respective items. The results presented strong evidence for the one-dimensionality of the testing materials as well as the distinct difficulty levels and, hence, convincing evidence for validity of the expected internal structure. Another argument for the interchangeability of the items within each of the six difficulty levels are the results of the randomly drawn item samples from the long version that were used for the short versions tested in Study 2. The two selected item samples had the highest and the lowest correlation with the PFT, respectively, based on the data in Study 1. Both item samples had five of 30 items in common and the two samples changed the order of the magnitude of their correlations with the PFT in Study 2. However, the results were limited, because in each of the three studies (Study 1 to 3) the sample size was too small to test LLTM conformity based on the likelihood tests. The fact, however, that all likelihood tests in all studies are not significant can be seen as cumulative evidence for the model conformity with the RMs and LLTMs. Furthermore, small sample sizes like in the present studies lead to low statistical power, but they deliver correct β -estimations (see Hohensinn et al., 2014; MacDonald & Kromrey, 2011). Further criteria such as the person separation reliability (PSR), the information criteria and the visual inspection of β -estimation plots provided additional evidence for model conformity. Hence, all criteria together strongly support the expected model fits for the R-Cube-Vis Test.

The small correlations with the grades in German and Mathematics were comparable with the small correlations between these criteria and the PFT and MRT in all studies. All studies confirmed that the stimulus materials were unbiased with respect to previous experience with Rubik's cubes.

The various aspects of evidence of validity strongly supported the R-Cube-Vis Test in its long and short version as a valid test for visualization that is distinguishable from spatial relations and can be placed at the corresponding end of the continuum of these two factors according to Pellegrino et al. (1984). Furthermore, the R-Cube-Vis Test fulfills the four requirements for the usage of eye tracking. The test by construction contains six homogenous item groups and is supported by its conformity with the LLTM. The items were created in such a way that all relevant visual features are separable and that the items are simply structured by showing only two figures. Finally, the test is computer-based. However, further research is advisable with larger samples from different populations and with alternative external criteria.

Eye Tracking measures as source of information of the R-Cube-Vis Test

In Study 4, eye tracking data were analyzed utilizing the short version of the R-Cube-Vis Test. At first, type- and item-specific characteristics of gaze patterns were disentangled to discriminate between those parts of the gaze patterns that were caused either by the demanded cognitive processing steps to solve the item or by visual item features that were independent of the necessary transformation steps. Secondly, the predictive power with respect to item accuracy was considered. The analyzed eye tracking measures were taken from four fixation-based Indicator-Classes. Two of these Indicator-Classes are based on fine-scaled Areas of Interest (AOIs), either on single cube features (Indicator-Class 1) or on fixations sequences on these features (Indicator-Class 4). The other two Indicator-Classes cluster more information together in large-scaled AOIs, either as whole cubes (Indicator-Class 2), or summing up fixations on these whole cubes over time as entropy (Indicator-Class 3). A general result of Study 4 showed descriptively that indicators using

large-scaled AOIs (Indicator-Class 2 and 3) remained stable over different items from the same homogenous group (type-specific) of the R-Cube-Vis Test, whereas indicators using fine-scaled AOIs (Indicator-Class 1 and 4) differed between items of the same item group (item-specific). This finding confirms the results of previous studies that also found stable results for large-scaled AOIs of complex items (e.g., Egan, 1979; Snow, 1980) and for small-scaled AOIs of simple AOIs (e.g., Just & Carpenter, 1976) but failed with small-scaled AOIs for complex items (e.g., Ivie & Embretson, 2010). However, there were differences between the difficulty levels. For example, the fixations on the three sides on the left cube seemed to be clearly item-specific in Level 4.1 but they descriptively showed type- as well as item-specific characteristics in Level 3.2.p and 4.2.c1. This is an important aspect if cognitive processing steps should be modeled for the R-Cube-Vis Test. If a gaze pattern model should describe processing steps of successful solving behavior in a specific level, it would be important to define a base model that reflects the type-specific characteristics found here. However, such a model should also include item-specific characteristics for each single item, maybe as an additive component. For example, the model of Level 4.1 would expect three fixations on the side that corresponds to the new color in general, but if the side is in the middle, the model would expect two fixations more. Future studies are necessary to gather more detailed information about each fine-scaled AOI and about its type-specific and item-specific behavior.

The Indicator-Class 4 contains the occurrence of bi-grams. Bi-grams are sequences of two consecutive fine-scaled fixations that are intended to indicate the cognitive processing steps that are necessary to solve the items. As the indicators of Indicator-Class 1 described above, the bi-grams showed item-specific gaze patterns. However, for each single item, the bi-grams were able to differentiate between correctly and incorrectly solved items. For example, the defined bi-grams were found more often for correctly solved items of Level 4.2.c1 and 4.2.c2 than for incorrectly solved items, although the absolute occurrence of these bi-grams differed between the items. Future studies might reveal further insights in the cognitive processes that are represented by the bi-grams. Thereby, bi-grams might be extended to n-grams with $n > 2$ that represent pre-defined and rationally motivated processing steps.

Such analyses using various fixation-based indicators should be extended with more items (e.g., using the long version of the R-Cube-Vis) to gain more information about the solving behavior applied to these items. In such a way, type- and item-specific fixation patterns of each homogenous item group can be identified. Type-specific characteristics of fixation patterns would describe the gaze patterns that are necessary to solve the item. The item-specific characteristics of fixation patterns could be included by using, e.g., an additive component specific for certain visual item features.

The second goal of Study 4 was to show the potential of fixation-based indicators to predict item accuracy. The number of fixations on the left cube, entropy values as well as bi-grams indicating necessary transformation steps seemed especially promising. However, differential results for the six difficulty levels were revealed. For example, the results suggested that for the levels of medium difficulty, focusing on the left cube seemed to be more expedient to solve the item, compared with the two most difficult levels. For the two most difficult levels, focusing on the right cube was related to a higher solving probability. A corresponding pattern existed for the entropy values. One reason might be that in both levels with medium difficulty it is harder to recognize the rotated elements by analyzing only the right cube because of the parallel rotation (especially in Level 3.2.p). Therefore, fixating the left cube provides more valid information to solve these items. In contrast, in the two most difficult levels, it is easier to see the rotated elements because of the crossed rotation. Hence, fixations on the right cube provide valid information to solve the item. However, follow-up studies are needed to confirm these explanations. In particular, the cognitive processing steps associated with the entropy values might be analyzed more thoroughly.

Limitations

Study 1 to 3 are limited with respect to the small sample size. On the one hand, this affects the interpretation of the correlations, which might deviate from the true relationship between the respective variables (see Schönbrodt & Perugini, 2013). On the other side, the statistical power is too low for the Andersen's likelihood

ratio test of the IRT models. However, even with small sample sizes the parameter estimation is robust (see Hohensinn et al., 2014; MacDonald & Kromrey, 2011). Still, the results in all three of the conducted studies were comparable, which supports the validity of the estimated correlations as well as the conformity of the tested IRT-models. The results of Study 4 are strongly explorative and have to be interpreted with caution. These results must be confirmed in follow-up studies.

Conclusion

The presented studies demonstrate the potential of eye tracking as a promising technique for revealing information about the cognitive processes applied for solving visual-spatial test items in the context of spatial thinking. Eye tracking measures provide information that goes beyond the information provided by accuracy and reaction times. However, eye tracking techniques can only be applied successfully if the stimulus materials are appropriate.

All considered fixation-based indicators differ between the six levels, not only in their absolute values, but also in their indication of correctly solved items. For example, higher entropy values of the medium difficulty levels predict a correctly solved item, whereas correctly solved items in the two most difficult levels are indicated by lower entropy values. Therefore, models of gaze patterns indicating specific cognitive processing steps have to be established separately for each of the difficulty levels.

If such gaze pattern models exist, they might be able to provide more detailed information about the cognitive steps that led to a certain answer. Based on information about how a participant attempted to solve an item, the success of solving the item can be predicted. This information can be utilized to improve the measurement in addition to accuracy and reaction times. On the one hand, such information could be used to create adaptive testing scenarios that include these gaze-related measures. On the other hand, it might even be possible to diagnose the participant's *potential* performance based on comparisons between the applied solving strategies and the optimal strategies. In this way, such a model could provide a differential diagnosis that could lead to a more sophisticated decision about a participant's ability.

Acknowledgement: This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation. Additionally, I want to thank all people who helped me to improve this article. Especially, I am very thankful for the many, very supportive discussions about the manuscript with Prof. Stefan Münzer who provided me a lot of valuable feedback to this work. Furthermore, I would like also to thank a lot my student assistant, Anna Klingauf, for her support in creating the stimulus materials, programming the experiments and conducting the studies. I am also very thankful for Maria Pesthy's proofreading of the manuscript. Last but not least, I want to thank the reviewers for their valuable remarks and all participants who attended the studies and solved (or at least tried to solve) all the Rubik's cube tasks of my tests.

Financial support: This research received no specific grant from any funding agency, commercial or nonprofit sectors.

Conflict of interests: The authors have no conflicts of interest to disclose.

Ethics statement: This research did not require ethical approval.

References

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (1999). *Intelligenz-Struktur-Test 2000 (I-S-T 2000)*: Göttingen: Hogrefe.
- Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38(1), 123–140. <https://doi.org/10.1007/BF02291180>
- Baghaei, P., & Kubinger, K. D. (2015). Linear Logistic Test Modeling with R. *Practical Assessment, Research & Evaluation*, 20(1), 1–11. Retrieved from <http://pareonline.net/getvn.asp?v=20&n=1>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA, USA: Addison-Wesley.
- Blajenkova, O., Kozhevnikov, M. [Maria], & Motes, M. A. (2006). Object-spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology*, 20(2), 239–263. <https://doi.org/10.1002/acp.1182>
- Bratko, D., Butkovic, A., Vukasovic, T., Chamorro-Premuzic, T., & Stumm, S. von (2012). Cognitive ability, self-assessed intelligence and personality: Common genetic but independent environmental aetiologies. *Intelligence*, 40(2), 91–99. <https://doi.org/10.1016/j.intell.2012.02.001>
- Brogden, H. E. (1946a). The effect of bias due to difficulty factors in product-moment item intercorrelations on the accuracy of estimation of reliability by the Kuder-Richardson formula number 20. *Educational and Psychological Measurement*, 6(4), 517–520. <https://doi.org/10.1177/001316444600600408>
- Brogden, H. E. (1946b). Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 11(4), 197–214. <https://doi.org/10.1007/BF02290130>
- Burnham, K. P., & Anderson, D. R. (2016). Multimodel inference. *Sociological Methods & Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Burton, L. J., & Fogarty, G. J. (2003). The factor structure of visual imagery and spatial abilities. *Intelligence*, 31(3), 289–318. [https://doi.org/10.1016/S0160-2896\(02\)00139-3](https://doi.org/10.1016/S0160-2896(02)00139-3)
- Carpenter, P. A. [P. A.], & Just, M. A. (1978). Eye fixations during mental rotation. In J. Senders, D. Fisher, & R. Monty (Eds.), *Eye movements and the higher psychological functions* (pp. 115–133). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511571312>
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, USA: Erlbaum.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One*, 10(3), e0121945. <https://doi.org/10.1371/journal.pone.0121945>
- Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., . . . Giannopoulos, I. (2018). The Index of Pupillary Activity. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *CHI 2018: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems : April 21-26, 2018, Montréal, QC, Canada* (pp. 1–13). New York, New York: The Association for Computing Machinery. <https://doi.org/10.1145/3173574.3173856>
- Egan, D. E. (1979). Testing based on understanding: Implications from studies of spatial ability. *Intelligence*, 3(1), 1–15. [https://doi.org/10.1016/0160-2896\(79\)80002-1](https://doi.org/10.1016/0160-2896(79)80002-1)
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976a). *Kit of Factor-Referenced Cognitive Tests*: Princeton, NJ: Educational Testing Service.
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976b). *Manual for kit of factor-referenced cognitive tests*: Princeton, NJ: Educational Testing Service.
- Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 24(2), 387–414. <https://doi.org/10.1037/0278-7393.24.2.387>
- Falissard, B. (2012). *psy: Various procedures used in psychometry*. R package version 1.1. Retrieved from <https://CRAN.R-project.org/package=psy>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fehringer, B. C. O. F. (2018). Improving the adaptive event detection algorithm of Nyström and Holmquist for noisy data. In L. Chuang, M. Burch, & K. Kurzhals (Eds.), *Proceedings of the 3rd Workshop on Eye Tracking and Visualization - ETVIS '18* (pp. 1–5). New York, New York, USA: ACM Press. <https://doi.org/10.1145/3205929.3205938>
- Fehringer, B. C. O. F. (2020). One threshold to rule them all? Modification of the Index of Pupillary Activity to optimize the indication of cognitive load. In A. Bulling, A. Huckauf, E. Jain, R. Radach, & D. Weiskopf (Eds.), *ACM Symposium on Eye Tracking Research and Applications* (pp. 1–5). New York, NY, USA: ACM. <https://doi.org/10.1145/3379156.3391341>
- Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, 14(1), 51–71. <https://doi.org/10.3102/00028312014001051>
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359–374. [https://doi.org/10.1016/0001-6918\(73\)90003-6](https://doi.org/10.1016/0001-6918(73)90003-6)

- Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, 77(4), 337–353. <https://doi.org/10.1006/jecp.2000.2594>
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, 36(3), 189–213. <https://doi.org/10.1177/0146621612439620>
- Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number line. *Developmental Psychology*, 48(5), 1229–1241. <https://doi.org/10.1037/a0027433>
- Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. *Perceptual and Motor Skills*, 43(3), 1319–1334. <https://doi.org/10.2466/pms.1976.43.3f.1319>
- Hegarty, M., & Kozhevnikov, M. [Maria] (1999). Types of visual-spatial representations and mathematical problem solving. *Journal of Educational Psychology*, 91(4), 684–689. <https://doi.org/10.1037/0022-0663.91.4.684>
- Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2), 175–191. <https://doi.org/10.1016/j.intell.2003.12.001>
- Hegarty, M., & Waller, D. A. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). Cambridge: Cambridge University Press.
- Höffler, T. N. (2010). Spatial Ability: Its Influence on Learning with Visualizations—a Meta-Analytic Review. *Educational Psychology Review*, 22(3), 245–269. <https://doi.org/10.1007/s10648-010-9126-7>
- Hohensinn, C., Kubinger, K. D., & Reif, M. (2014). On robustness and power of the likelihood-ratio test as a model of the linear logistic test model. *Journal of Applied Measurement*, 15(3), 252–266.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford: OUP Oxford. Retrieved from <http://gbv.eblib.com/patron/FullRecord.aspx?p=990529>
- Ivie, J. L., & Embretson, S. E. (2010). Cognitive process modeling of spatial ability: The assembling objects task. *Intelligence*, 38(3), 324–335. <https://doi.org/10.1016/j.intell.2010.02.002>
- Jacobs, K. E., & Roodenburg, J. (2014). The development and validation of the self-report measure of cognitive abilities: A multitrait-multimethod study. *Intelligence*, 42, 5–21. <https://doi.org/10.1016/j.intell.2013.09.004>
- Jansen, P., Schmelter, A., Quaiser-Pohl, C., Neuburger, S., & Heil, M. [M.] (2013). Mental rotation performance in primary school age children: Are there gender differences in chronometric tests? *Cognitive Development*, 28(1), 51–62. <https://doi.org/10.1016/j.cogdev.2012.08.005>
- Jansen-Osmann, P., & Heil, M. [Martin] (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain and Cognition*, 64(3), 217–227. <https://doi.org/10.1016/j.bandc.2007.03.002>
- Jones, E., Oliphant, T., Peterson, P., & others (2001). *SciPy: Open source scientific tools for Python*. Retrieved from <http://www.scipy.org/>
- Just, M. A., & Carpenter, P. A. [Patricia A.] (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4), 441–480. [https://doi.org/10.1016/0010-0285\(76\)90015-3](https://doi.org/10.1016/0010-0285(76)90015-3)
- Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, 35(3), 211–223. <https://doi.org/10.1016/j.intell.2006.07.009>
- Kozhevnikov, M. [Maria], & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29(5), 745–756. <https://doi.org/10.3758/BF03200477>
- Kozhevnikov, M. [Maria], Hegarty, M., & Mayer, R. E. (2002). Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. *Cognition and Instruction*, 20(1), 47–77. https://doi.org/10.1207/S1532690XCI2001_3
- Kozhevnikov, M. [Maria], Kozhevnikov, M. [Michael], Yu, C. J., & Blazhenkova, O. (2013). Creativity, visualization abilities, and visual cognitive style. *The British Journal of Educational Psychology*, 83(Pt 2), 196–209. <https://doi.org/10.1111/bjep.12013>
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160. <https://doi.org/10.1007/BF02288391>
- Kyllonen, P. C., Lohman, D. F., & Snow, R. E. [Richard E.] (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology*, 76(1), 130–145. <https://doi.org/10.1037/0022-0663.76.1.130>
- Kyllonen, P. C., Lohman, D. F., & Woltz, D. J. (1984). Componential modeling of alternative strategies for performing spatial tasks. *Journal of Educational Psychology*, 76(6), 1325–1345. <https://doi.org/10.1037/0022-0663.76.6.1325>
- LeFevre, J.-A., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, 81(6), 1753–1767. <https://doi.org/10.1111/j.1467-8624.2010.01508.x>
- Likert, R., & Quasha, W. H. (1970). *Manual for the Revised Minnesota Paper Form Board Test*. New York: The Psychological Corporation.

- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 181–248). Hillsdale, NJ, England: Lawrence Erlbaum Associates.
- MacDonald, G., & Kromrey, J. (2011). Linear logistic test model: Using SAS® to simulate the decomposition of item difficulty by algorithm, sample size, cognitive component and time to convergence. In American Statistical Association (Chair), *Proceedings of the American Statistical Association's Joint Statistical Meeting, Social Statistics Section*, Miami, FL.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A user's guide* (Second edition). Hove: Psychology Press.
- Mair, P., Hatzinger, R., & Maier M. J. (2015). *eRm: Extended Rasch modeling*. 0.15-6. Retrieved from <http://erm.r-forge.r-project.org/>
- McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86(5), 889–918. <https://doi.org/10.1037/0033-2909.86.5.889>
- McKinney, W. (2011). pandas: A Foundational Python Library for Data Analysis and Statistics.
- Meißner, A., Greiff, S., Frischkorn, G. T., & Steinmayr, R. (2016). Predicting complex problem solving and school grades with working memory and ability self-concept. *Learning and Individual Differences*, 49, 323–331. <https://doi.org/10.1016/j.lindif.2016.04.006>
- Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. (1957). The description of spatial-visualization abilities. *Educational and Psychological Measurement*, 17(2), 185–199. <https://doi.org/10.1177/001316445701700202>
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related?: A latent-variable analysis. *Journal of Experimental Psychology: General*, 130(4), 621–640. <https://doi.org/10.1037/0096-3445.130.4.621>
- Münzer, S. (2012). Facilitating spatial perspective taking through animation: Evidence from an aptitude–treatment–interaction. *Learning and Individual Differences*, 22(4), 505–510. <https://doi.org/10.1016/j.lindif.2012.03.002>
- Münzer, S. (2015). Facilitating recognition of spatial structures through animation and the role of mental rotation ability. *Learning and Individual Differences*, 38, 76–82. <https://doi.org/10.1016/j.lindif.2014.12.007>
- Nazareth, A. (2015). *Factors affecting adult mental rotation performance*. (Doctoral dissertation). Retrieved from <http://digital-commons.fiu.edu/etd/>
- Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1), 188–204. <https://doi.org/10.3758/BRM.42.1.188>
- Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence*, 47, 141–158. <https://doi.org/10.1016/j.intell.2014.09.002>
- Pässler, K., Beinicke, A., & Hell, B. (2015). Interests and intelligence: A meta-analysis. *Intelligence*, 50, 30–51. <https://doi.org/10.1016/j.intell.2015.02.001>
- Pellegrino, J. W., Alderton, D. L., & Shute, V. J. (1984). Understanding spatial ability. *Educational Psychologist*, 19(3), 239–253. <https://doi.org/10.1080/00461528409529300>
- Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition*, 28(1), 39–58. <https://doi.org/10.1006/brcg.1995.1032>
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3–25. <https://doi.org/10.1080/00335558008248231>
- Psychology Software Tools Inc. (2011). *E-Prime 2.0: Extension for tobi*. Pittsburgh, PA: Psychology Software Tools.
- Psychology Software Tools Inc. (2012). *E-Prime 2.0*. Retrieved from <http://www.pstnet.com>
- R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163. <https://doi.org/10.2307/271063>
- Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois. Retrieved from <http://CRAN.R-project.org/package=psych>
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software*, 17(5), 1–25. Retrieved from <http://www.jstatsoft.org/v17/i05/>
- Royston, P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Applied Statistics*, 44(4), 547–551. <https://doi.org/10.2307/2986146>
- Sakamoto, Y., Ishiguro, M., & Kitagawa G. (1986). *Akaike Information Criterion statistics*. New York, NY, USA: D. Reidel Publishing Company.
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA)*.
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

- Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, 24, 42–52. <https://doi.org/10.1016/j.lindif.2012.12.011>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701–703.
- Snow, R. E. [R. E.] (1980). Aptitude processes. In R. E. Snow, P. A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction (Vol. 1): Cognitive process analyses of aptitude*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates.
- Sun, W., Chou, C.-P., Stacy, A. W., Ma, H., Unger, J., & Gallaher, P. (2007). SAS and SPSS macros to calculate standardized Cronbach's alpha using the upper bound of the phi coefficient for dichotomous items. *Behavior Research Methods*, 39(1), 71–81. <https://doi.org/10.3758/BF03192845>
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1(1), 1–26. <https://doi.org/10.1111/1529-1006.001>
- Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, 135(2), 77–99. <https://doi.org/10.1016/j.actpsy.2010.02.006>
- Thurstone, T. G. (1958). *Manual for the SRA primary mental abilities*: Chicago, Science Research Associates.
- Uzzaman, S., & Joordens, S. (2011). The eyes know what you are thinking: Eye movements as an objective measure of mind wandering. *Consciousness and Cognition*, 20(4), 1882–1886. <https://doi.org/10.1016/j.concog.2011.09.010>
- Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- Van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22–30. <https://doi.org/10.1109/MCSE.2011.37>
- Van Rossum, G. (1995, January 1). *Python Tutorial, Technical Report CS-R9526*. Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47(2), 599–604. <https://doi.org/10.2466/pms.1978.47.2.599>
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817–835. <https://doi.org/10.1037/a0016127>
- Wang, L., & Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educational Psychologist*, 49(4), 261–282. <https://doi.org/10.1080/00461520.2014.960568>
- Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12), 1–20. Retrieved from <http://www.jstatsoft.org/v21/i12/>
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*: Springer New York. Retrieved from <http://had.co.nz/ggplot2/book>
- Wickham, H. (2011). The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1–29. Retrieved from <http://www.jstatsoft.org/v40/i01/>
- Wickham, H., François, R., Henry, L., & Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, US-DE: Wide Range, Inc. Retrieved from <https://www.rasch.org/measess/me-all.pdf>
- Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14(6), 1–27. <https://doi.org/10.18637/jss.v014.i06>

Appendix

Study 1

Table 22: Item statistics of the Rasch models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.1													
4.1_1	50	1.50	1.07	2.90	1.62	1.09	4.63	1.29	1.03	1.28	2.06	0.91	0.45
4.1_2	50							2.03	0.95	0.31	0.97	0.73	0.44
4.1_3	50										1.58	0.87	0.46
4.1_4	50				1.62	0.93	0.25	2.03	0.99	0.45	1.58	1.16	0.66
4.1_5	50										1.81	1.45	1.83
4.1_6	50							2.03	0.95	0.31	2.06	0.97	0.58
4.1_7	50							2.03	0.90	0.24	2.06	1.13	0.57
4.1_8	50										3.05	1.29	1.11
4.1_9	50	1.50	1.06	1.55	1.62	0.92	0.23	1.29	0.95	0.37	1.58	1.13	0.79
4.1_10	50										1.16	0.71	0.42
4.1_11	50	1.50	1.06	1.78	0.87	0.98	1.40	1.29	1.00	2.72	1.58	1.02	1.64
4.1_12	50										2.34	1.25	0.82
4.1_13	50										2.34	1.01	0.51
4.1_14	50	1.50	1.05	1.07	1.62	1.07	1.56	2.03	1.04	1.09	1.81	1.05	0.58
4.1_15	50	0.76	1.11	0.98	0.87	1.04	0.82	0.50	1.00	0.93	1.36	1.12	1.79
4.1_16	50										2.06	1.27	0.65
4.1_17	50							1.29	1.03	0.62	1.58	1.25	0.76
4.1_18	50	1.50	1.04	0.86	1.62	1.06	1.22	2.03	1.04	1.20	1.36	1.12	0.80
4.1_19	50	1.50	0.79	0.14	1.62	0.84	0.16	2.03	0.87	0.20	1.58	1.14	0.71
4.1_20	50							2.03	0.95	0.31	2.06	1.08	0.77
4.1_21	50							2.03	0.95	0.31	1.81	0.84	0.44
4.1_22	50	0.76	1.14	1.74	0.87	1.18	2.51	1.29	1.00	0.93	2.06	0.79	0.41
4.1_23	50	1.50	0.97	0.35	1.62	0.98	0.35	1.29	0.96	0.40	1.58	0.81	0.50
4.1_24	50										1.81	1.54	1.29
Level 3.1													
3.1_1	50	0.76	1.02	0.60	0.87	1.06	0.77	1.29	1.00	0.64	1.81	0.59	0.26
3.1_2	50							1.29	1.07	1.34	0.62	0.91	1.03
3.1_3	50	1.50	0.97	0.35	1.62	0.98	0.35	2.03	0.95	0.31	1.36	0.89	0.48
3.1_4	50							1.29	1.01	2.35	0.62	1.14	1.15
3.1_5	50										0.80	0.91	0.56
3.1_6	50	1.50	1.07	2.90	1.62	1.09	4.63	2.03	1.06	2.32	0.80	1.05	1.46
3.1_7	50	0.76	0.88	0.93	0.87	0.94	0.93	1.29	0.93	0.51	-0.14	0.81	0.86
3.1_8	50				1.62	1.05	0.97	0.83	1.11	0.94	0.62	1.15	1.05
3.1_9	50							0.83	1.04	0.99	0.62	0.98	1.17
3.1_10	50	0.76	1.17	4.83	0.87	1.19	3.73	1.29	1.01	2.35	0.80	1.16	0.91

Table 22 (continued): Item statistics of the Rasch models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N
Level 3.1													
3.1_11	50	1.50	0.97	0.35	1.62	0.98	0.35	1.29	0.89	0.29	0.97	0.97	1.65
3.1_12	50	1.50	0.97	0.35	1.62	0.98	0.35	1.29	0.96	0.40	0.97	1.14	1.11
3.1_13	50	1.50	1.08	26.34	1.62	1.09	44.94	1.29	1.04	4.65	1.58	0.79	1.74
3.1_14	50	0.76	1.12	1.32	0.87	1.16	1.50	1.29	1.10	1.07	0.80	1.20	0.91
3.1_15	50	1.50	0.97	0.35	1.62	0.98	0.35	2.03	0.95	0.31	1.81	1.05	0.60
3.1_16	50	0.76	0.73	0.19	0.87	0.74	0.18	0.50	0.78	0.32	0.80	0.99	0.80
3.1_17	50	1.50	1.05	1.07	1.62	0.96	0.31	0.50	0.98	1.15	1.36	0.86	0.71
3.1_18	50	1.50	1.04	0.86	1.62	1.06	1.22	0.83	1.00	0.63	1.36	1.07	0.90
3.1_19	50	1.50	0.79	0.14	1.62	0.84	0.16	0.22	1.05	1.30	0.30	1.28	1.52
3.1_20	50							2.03	0.95	0.31	1.81	0.76	0.40
3.1_21	50				1.62	0.98	0.35	0.50	0.89	0.39	0.80	1.18	0.96
3.1_22	50	1.50	1.04	0.86	1.62	1.06	1.22	0.83	1.12	1.18	1.16	0.63	0.66
3.1_23	50	1.50	1.04	0.86	1.62	1.06	1.22	1.29	0.99	0.72	1.36	1.39	0.87
3.1_24	50										0.97	1.04	0.86
Level 4.2.p													
4.2_1	50	-0.31	0.83	0.51	-0.22	0.80	0.43	-0.01	0.85	0.56	-0.14	0.91	0.94
4.2_2	49	0.71	0.89	1.71	0.83	0.96	2.73	0.78	1.07	3.53	0.37	0.95	1.18
4.2_3	50	1.50	1.04	0.91	1.62	1.00	0.45	2.03	1.02	0.76	0.46	0.80	0.60
4.2_4	49				1.58	0.87	0.19	1.98	0.90	0.23	0.06	1.07	1.04
4.2_5	50	0.76	0.79	0.25	0.87	0.84	0.26	0.83	0.84	0.32	0.46	0.84	0.65
4.2_6	49	-0.09	0.74	0.91	0.01	0.80	1.35	0.44	0.91	2.20	-0.24	1.15	1.22
4.2_7	50	0.76	0.98	0.64	0.87	0.97	0.85	0.83	0.98	0.59	0.80	1.01	1.26
4.2_8	50	0.76	0.88	0.93	0.87	0.93	0.74	0.50	0.87	0.67	0.15	0.89	0.69
4.2_9	50	0.76	0.73	0.19	0.87	0.74	0.18	0.83	0.82	0.32	0.00	1.07	0.89
4.2_10	49	0.71	0.88	0.90	0.83	0.95	1.36	1.24	0.97	2.62	0.71	0.93	1.11
4.2_11	50	0.76	0.92	0.34	0.87	0.91	0.36	0.83	0.87	0.33	0.80	0.87	0.55
4.2_12	50	1.50	0.79	0.14	1.62	0.84	0.16	0.83	0.95	0.58	0.97	1.07	0.85
4.2_13	50	0.30	0.94	0.61	0.41	0.97	0.81	0.50	1.08	1.18	0.30	0.72	0.71
4.2_14	50	1.50	1.07	3.53	1.62	1.08	2.35	2.03	0.98	0.42	1.36	0.89	0.78
4.2_15	50	0.30	0.84	0.49	0.41	0.81	0.28	0.22	0.86	0.59	0.00	0.80	0.93
4.2_16	50	0.30	0.79	0.40	0.41	0.83	0.50	0.22	0.90	0.69	0.30	0.94	1.00
4.2_17	50	0.76	0.73	0.19	0.87	0.74	0.18	1.29	0.81	0.23	0.15	0.65	0.43
4.2_18	50	0.30	1.03	0.81	0.41	1.06	1.11	0.83	1.06	1.43	0.15	0.94	0.89
4.2_19	49	0.25	0.95	0.56	0.36	0.96	0.72	-0.07	0.98	0.66	0.21	0.78	0.53
4.2_20	49							0.44	0.91	0.55	-0.09	0.59	0.39
4.2_21	50	1.50	0.97	0.35	1.62	0.98	0.35	1.29	0.95	0.37	0.46	0.81	0.54
4.2_22	50	0.76	0.90	2.83	0.87	0.96	4.62	1.29	0.97	3.18	1.16	0.74	0.48
4.2_23	50	0.76	0.96	0.37	0.87	1.01	0.44	0.22	0.87	0.41	0.46	0.97	1.14
4.2_24	50	1.50	0.79	0.14	0.87	0.74	0.18	0.50	0.77	0.31	-0.14	0.87	0.71

Table 22 (continued): Item statistics of the Rasch models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2					wACC: split between 2 and 3					wACC: split between 3 and 4				
Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N
Level 3.2.p																			
3.2_1	50	-0.03	0.90	0.78	-0.22	1.06	1.00	-0.40	0.87	0.62	-0.55	0.92	0.69						
3.2_2	50	-0.03	1.23	1.32	0.06	1.16	1.46	-0.01	1.10	1.56	-0.55	1.21	1.09						
3.2_3	50	-0.54	0.76	0.47	-0.46	0.85	0.52	-0.73	0.96	0.67	-0.82	1.04	0.84						
3.2_4	50	-0.31	0.90	0.75	-0.22	1.03	1.04	-0.40	0.98	1.09	-0.14	0.94	0.92						
3.2_5	50	-0.54	1.06	1.71	-0.46	1.16	1.95	-0.40	1.20	1.70	-0.55	1.35	1.29						
3.2_6	50	0.30	0.95	0.83	0.41	0.98	1.16	-0.22	0.98	1.00	-0.94	0.97	0.77						
3.2_7	50	0.30	0.80	0.61	0.41	0.73	0.22	0.50	0.90	0.41	-0.68	1.35	1.58						
3.2_8	50	-0.03	0.75	0.38	0.06	0.81	0.46	-0.40	1.17	1.82	-0.14	1.09	1.22						
3.2_9	50	1.50	1.04	0.91	1.62	1.00	0.45	0.50	0.95	0.47	-0.42	0.87	0.77						
3.2_10	50	-0.54	0.94	0.74	-0.86	0.91	0.60	-0.73	0.98	0.77	-1.33	1.06	0.97						
3.2_11	50	-0.03	0.97	0.69	-0.46	1.05	0.89	-0.40	0.98	0.66	-0.14	1.24	1.39						
3.2_12	50	-0.74	1.05	1.35	-0.67	0.96	0.84	-0.57	1.02	0.88	-0.28	1.23	1.33						
3.2_13	50	-0.03	0.85	0.61	0.06	0.96	0.83	-0.73	1.00	0.79	-0.68	1.12	1.22						
3.2_14	50	-0.03	0.99	0.82	-0.46	0.97	0.79	-0.40	1.02	0.78	-0.42	1.04	1.23						
3.2_15	50	-0.03	1.03	0.83	0.06	0.97	0.70	-0.22	1.08	0.82	-0.55	0.95	0.97						
3.2_16	50	-0.03	0.92	0.58	-0.46	0.94	0.74	-0.40	0.89	0.51	-0.14	0.83	0.61						
3.2_17	50	1.50	0.90	0.22	1.62	0.93	0.25	-0.01	1.08	0.71	-0.94	0.86	0.67						
3.2_18	50	-0.03	0.96	0.90	0.06	0.90	0.63	-0.40	0.88	0.61	-0.94	1.24	1.17						
3.2_19	50	0.76	0.98	0.86	0.87	0.86	0.25	-0.40	0.77	0.41	-0.28	0.91	0.70						
3.2_20	50	1.50	0.79	0.14	1.62	0.84	0.16	1.29	0.92	0.48	-0.82	0.90	0.87						
3.2_21	50	1.50	1.07	2.90	1.62	1.09	4.63	1.29	0.93	0.38	-0.42	1.08	0.82						
3.2_22	50	-0.03	0.94	1.22	0.06	1.05	1.83	-0.01	1.00	1.10	-0.55	0.92	0.72						
3.2_23	50	-0.74	1.02	1.26	-0.86	0.91	1.28	-0.73	1.04	1.91	-0.68	1.14	1.92						
3.2_24	50	-0.54	1.01	0.82	-0.67	1.00	0.78	-0.57	1.01	0.76	-0.68	1.10	1.49						
Level 4.2.c1																			
4.3_1	49	-1.41	0.83	0.78	-1.65	0.76	0.78	-1.46	0.99	1.19	-1.57	1.39	1.44						
4.3_2	50	-1.87	0.89	0.77	-1.99	0.85	0.69	-1.90	1.03	0.95	-1.20	1.08	0.95						
4.3_5	48	-1.36	0.83	0.62	-1.66	0.78	0.62	-2.06	0.92	0.86	-2.07	1.01	2.08						
4.3_7	48	-0.80	1.14	1.24	-0.95	1.08	1.37	-1.12	1.16	1.67	-1.12	1.29	1.57						
4.3_9	50	-1.63	0.99	1.01	-1.99	0.95	0.97	-2.34	1.14	1.20	-1.59	0.91	0.91						
4.3_10	49	-1.17	0.99	0.89	-1.45	0.91	0.91	-1.64	1.01	1.04	-1.70	0.86	1.50						
4.3_13	48	-1.64	0.82	0.64	-1.92	0.78	0.65	-1.95	0.72	0.61	-1.65	0.76	0.58						
4.3_15	49	-1.72	1.10	1.17	-2.08	1.01	1.01	-2.22	0.91	0.99	-2.10	1.05	1.45						
4.3_17	47	-2.05	1.07	1.13	-2.44	1.02	1.10	-2.59	1.23	1.45	-2.34	1.28	1.26						
4.3_18	48	-1.36	1.00	0.79	-1.62	0.93	0.79	-1.93	0.87	0.74	-1.69	0.82	0.63						
4.3_21	48	-1.15	0.92	0.82	-1.45	0.87	0.85	-1.39	0.87	0.70	-1.39	0.73	0.56						
4.3_23	48	-2.00	1.31	1.44	-2.39	1.24	1.43	-2.52	1.23	1.57	-2.07	1.15	1.13						
Level 4.2.c2																			
4.3_3	49	-3.05	1.14	1.40	-3.43	1.02	1.26	-3.44	0.98	1.26	-3.37	0.78	0.64						

Table 22 (continued): Item statistics of the Rasch models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N	Difficulty	Infit	Outfit	Difficulty	Item	N
Level 4.2.c2													
4.3_4	50	-3.26	0.93	0.98	-3.55	0.90	1.11	-3.69	1.00	1.43	-4.35	0.91	2.45
4.3_6	49	-2.96	0.87	0.86	-3.65	0.92	0.97	-4.21	0.80	0.61	-3.80	0.98	0.57
4.3_8	49	-2.67	0.86	0.91	-3.04	0.81	0.93	-3.56	0.86	1.29	-3.00	0.88	0.64
4.3_11	49	-2.27	1.11	1.14	-2.63	1.17	1.20	-3.21	1.07	1.07	-2.68	0.77	0.58
4.3_12	48	-2.83	0.94	0.89	-3.54	0.99	1.00	-4.20	1.05	1.57	-4.04	0.74	2.15
4.3_14	49	-2.84	1.31	1.59	-3.54	1.18	1.57	-3.67	1.08	1.14	-3.56	0.65	0.36
4.3_16	48	-2.99	0.97	0.99	-3.59	0.88	0.92	-3.61	0.97	1.10	-3.30	0.91	0.65
4.3_19	48	-2.63	1.15	1.27	-3.02	1.11	1.28	-3.31	1.02	0.98	-2.81	0.95	0.91
4.3_20	48	-2.83	1.12	1.31	-3.22	1.15	1.47	-4.20	0.75	0.64	-4.04	0.89	0.84
4.3_22	49	-2.84	1.13	1.30	-3.52	1.12	1.70	-3.92	1.26	2.66	-3.37	1.27	2.71
4.3_24	50	-2.29	0.97	1.02	-3.04	0.96	0.91	-3.56	1.03	0.88	-3.57	1.16	0.66

Table 23: Item statistics of the linear logistic test models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.1													
4.1_1	50	0.55	1.18	3.14	0.65	1.26	5.26	0.85	1.61	2.06	0.86	0.81	0.41
4.1_2	50							0.85	0.78	0.27	0.86	1.16	0.70
4.1_3	50										0.86	0.97	0.50
4.1_4	50				0.65	1.11	0.30	0.85	0.82	0.38	0.86	1.24	0.73
4.1_5	50										0.86	1.41	1.73
4.1_6	50							0.85	0.78	0.27	0.86	0.88	0.51
4.1_7	50							0.85	0.75	0.21	0.86	1.01	0.51
4.1_8	50										0.86	1.00	0.54
4.1_9	50	0.55	1.17	1.69	0.65	1.10	0.28	0.85	1.54	0.59	0.86	1.21	0.87
4.1_10	50										0.86	1.02	0.57
4.1_11	50	0.55	1.17	1.94	0.65	2.27	3.35	0.85	1.59	4.34	0.86	1.09	1.86
4.1_12	50										0.86	1.04	0.62
4.1_13	50										0.86	0.82	0.42
4.1_14	50	0.55	1.15	1.17	0.65	1.24	1.80	0.85	0.87	0.88	0.86	1.03	0.57
4.1_15	50	0.55	2.25	2.20	0.65	2.33	1.94	0.85	3.10	3.13	0.86	1.28	2.44
4.1_16	50										0.86	1.13	0.58
4.1_17	50							0.85	1.62	1.01	0.86	1.32	0.84
4.1_18	50	0.55	1.14	0.94	0.65	1.24	1.42	0.85	0.88	0.96	0.86	1.29	1.04
4.1_19	50	0.55	0.91	0.16	0.65	1.02	0.19	0.85	0.71	0.18	0.86	1.21	0.79
4.1_20	50							0.85	0.78	0.27	0.86	0.98	0.65

Table 23 (continued): Item statistics of the linear logistic test models for ACC and all splits of wACC

ACC					wACC:			wACC:			wACC:		
					split between 1 and 2			split between 2 and 3			split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.1													
4.1_21	50							0.85	0.78	0.27	0.86	0.83	0.44
4.1_22	50	0.55	2.29	3.92	0.65	2.46	6.00	0.85	1.58	1.50	0.86	0.70	0.36
4.1_23	50	0.55	1.08	0.40	0.65	1.16	0.42	0.85	1.55	0.64	0.86	0.89	0.56
4.1_24	50										0.86	1.49	1.24
Level 3.1													
3.1_1	50	0.00	1.32	0.79	0.00	1.32	0.98	0.00	0.79	0.49	0.00	0.50	0.29
3.1_2	50							0.00	0.85	0.98	0.00	0.98	1.22
3.1_3	50	0.00	0.67	0.25	0.00	0.66	0.25	0.00	0.42	0.16	0.00	0.78	0.45
3.1_4	50							0.00	0.80	1.65	0.00	1.23	1.29
3.1_5	50										0.00	0.94	0.58
3.1_6	50	0.00	0.77	1.84	0.00	0.77	2.78	0.00	0.53	0.83	0.00	1.06	1.55
3.1_7	50	0.00	1.18	1.22	0.00	1.20	1.19	0.00	0.72	0.40	0.00	1.39	1.11
3.1_8	50				0.00	0.74	0.62	0.00	1.21	1.04	0.00	1.23	1.16
3.1_9	50							0.00	1.14	1.09	0.00	1.05	1.40
3.1_10	50	0.00	1.46	6.32	0.00	1.45	4.70	0.00	0.80	1.65	0.00	1.18	0.94
3.1_11	50	0.00	0.67	0.25	0.00	0.66	0.25	0.00	0.68	0.24	0.00	0.94	1.51
3.1_12	50	0.00	0.67	0.25	0.00	0.66	0.25	0.00	0.74	0.32	0.00	1.10	1.05
3.1_13	50	0.00	0.78	16.31	0.00	0.78	26.21	0.00	0.83	3.20	0.00	0.73	1.03
3.1_14	50	0.00	1.42	1.73	0.00	1.42	1.89	0.00	0.88	0.78	0.00	1.21	0.96
3.1_15	50	0.00	0.67	0.25	0.00	0.66	0.25	0.00	0.42	0.16	0.00	0.87	0.55
3.1_16	50	0.00	1.04	0.25	0.00	1.02	0.23	0.00	1.22	0.45	0.00	1.01	0.83
3.1_17	50	0.00	0.74	0.70	0.00	0.65	0.22	0.00	1.41	1.70	0.00	0.77	0.59
3.1_18	50	0.00	0.73	0.57	0.00	0.75	0.77	0.00	1.10	0.70	0.00	0.94	0.75
3.1_19	50	0.00	0.50	0.11	0.00	0.54	0.12	0.00	1.80	2.48	0.00	1.46	2.31
3.1_20	50							0.00	0.42	0.16	0.00	0.65	0.38
3.1_21	50				0.00	0.66	0.25	0.00	1.33	0.55	0.00	1.21	0.98
3.1_22	50	0.00	0.73	0.57	0.00	0.75	0.77	0.00	1.22	1.30	0.00	0.59	0.57
3.1_23	50	0.00	0.73	0.57	0.00	0.75	0.77	0.00	0.77	0.54	0.00	1.21	0.78
3.1_24	50										0.00	1.00	0.82
Level 4.2.p													
4.2_1	50	-0.51	1.79	1.03	-0.51	1.71	0.85	-0.33	1.41	0.90	-0.60	1.00	1.22
4.2_2	49	-0.51	0.80	1.45	-0.51	0.84	2.18	-0.33	0.97	2.90	-0.60	0.92	1.10
4.2_3	50	-0.51	0.54	0.40	-0.51	0.50	0.22	-0.33	0.44	0.27	-0.60	0.77	0.57
4.2_4	49				-0.51	0.40	0.12	-0.33	0.34	0.13	-0.60	1.11	1.16
4.2_5	50	-0.51	0.66	0.22	-0.51	0.69	0.23	-0.33	0.72	0.29	-0.60	0.81	0.62
4.2_6	49	-0.51	1.39	1.61	-0.51	1.40	2.32	-0.33	1.05	2.50	-0.60	1.29	1.62
4.2_7	50	-0.51	0.83	0.54	-0.51	0.82	0.68	-0.33	0.85	0.51	-0.60	0.91	1.02
4.2_8	50	-0.51	0.75	0.76	-0.51	0.78	0.60	-0.33	0.97	0.73	-0.60	0.91	0.72
4.2_9	50	-0.51	0.61	0.18	-0.51	0.61	0.17	-0.33	0.70	0.29	-0.60	1.14	0.97

Table 23 (continued): Item statistics of the linear logistic test models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.2.p													
4.2_10	49	-0.51	0.78	0.78	-0.51	0.83	1.10	-0.33	0.65	1.41	-0.60	0.87	0.88
4.2_11	50	-0.51	0.78	0.30	-0.51	0.76	0.31	-0.33	0.74	0.30	-0.60	0.77	0.54
4.2_12	50	-0.51	0.32	0.10	-0.51	0.36	0.11	-0.33	0.82	0.50	-0.60	0.94	0.79
4.2_13	50	-0.51	1.17	0.76	-0.51	1.16	0.97	-0.33	1.17	1.29	-0.60	0.72	0.70
4.2_14	50	-0.51	0.57	1.38	-0.51	0.58	0.90	-0.33	0.40	0.18	-0.60	0.83	0.63
4.2_15	50	-0.51	1.06	0.61	-0.51	1.01	0.34	-0.33	1.19	0.81	-0.60	0.85	1.12
4.2_16	50	-0.51	1.02	0.50	-0.51	1.03	0.61	-0.33	1.23	0.94	-0.60	0.93	0.98
4.2_17	50	-0.51	0.61	0.18	-0.51	0.61	0.17	-0.33	0.47	0.17	-0.60	0.68	0.46
4.2_18	50	-0.51	1.26	1.00	-0.51	1.25	1.33	-0.33	0.93	1.15	-0.60	0.97	0.94
4.2_19	49	-0.51	1.23	0.73	-0.51	1.20	0.90	-0.33	1.62	1.13	-0.60	0.80	0.54
4.2_20	49							-0.33	1.05	0.64	-0.60	0.70	0.46
4.2_21	50	-0.51	0.47	0.18	-0.51	0.48	0.18	-0.33	0.59	0.25	-0.60	0.77	0.52
4.2_22	50	-0.51	0.77	2.26	-0.51	0.81	3.48	-0.33	0.63	1.62	-0.60	0.68	0.47
4.2_23	50	-0.51	0.81	0.33	-0.51	0.85	0.37	-0.33	1.21	0.54	-0.60	0.92	1.05
4.2_24	50	-0.51	0.32	0.10	-0.51	0.61	0.17	-0.33	0.87	0.34	-0.60	0.99	0.84
Level 3.2.p													
3.2_1	50	-1.05	0.94	0.80	-1.16	1.19	1.16	-1.18	0.96	0.68	-1.46	0.90	0.68
3.2_2	50	-1.05	1.25	1.35	-1.16	1.10	1.33	-1.18	0.99	1.27	-1.46	1.19	1.07
3.2_3	50	-1.05	1.22	0.70	-1.16	1.18	0.72	-1.18	1.27	0.90	-1.46	1.05	0.89
3.2_4	50	-1.05	1.14	0.96	-1.16	1.16	1.20	-1.18	1.06	1.21	-1.46	0.92	0.81
3.2_5	50	-1.05	1.53	2.74	-1.16	1.48	2.81	-1.18	1.28	1.90	-1.46	1.32	1.26
3.2_6	50	-1.05	0.77	0.65	-1.16	0.75	0.80	-1.18	0.97	0.97	-1.46	1.01	0.85
3.2_7	50	-1.05	0.64	0.48	-1.16	0.52	0.20	-1.18	0.61	0.32	-1.46	1.35	1.49
3.2_8	50	-1.05	0.79	0.40	-1.16	0.78	0.44	-1.18	1.25	2.02	-1.46	1.05	1.13
3.2_9	50	-1.05	0.44	0.29	-1.16	0.42	0.19	-1.18	0.65	0.35	-1.46	0.86	0.73
3.2_10	50	-1.05	1.40	1.14	-1.16	1.65	1.09	-1.18	1.27	1.06	-1.46	1.22	1.17
3.2_11	50	-1.05	1.00	0.71	-1.16	1.38	1.25	-1.18	1.07	0.72	-1.46	1.20	1.21
3.2_12	50	-1.05	1.73	2.56	-1.16	1.48	1.39	-1.18	1.21	1.10	-1.46	1.18	1.25
3.2_13	50	-1.05	0.88	0.63	-1.16	0.92	0.77	-1.18	1.30	1.08	-1.46	1.12	1.18
3.2_14	50	-1.05	1.02	0.84	-1.16	1.29	1.11	-1.18	1.10	0.86	-1.46	1.01	1.21
3.2_15	50	-1.05	1.06	0.86	-1.16	0.92	0.66	-1.18	1.06	0.80	-1.46	0.93	0.94
3.2_16	50	-1.05	0.96	0.60	-1.16	1.27	1.04	-1.18	0.99	0.56	-1.46	0.78	0.63
3.2_17	50	-1.05	0.32	0.13	-1.16	0.36	0.15	-1.18	0.96	0.62	-1.46	0.92	0.70
3.2_18	50	-1.05	0.99	0.92	-1.16	0.85	0.59	-1.18	0.98	0.67	-1.46	1.28	1.20
3.2_19	50	-1.05	0.59	0.47	-1.16	0.46	0.18	-1.18	0.87	0.45	-1.46	0.87	0.70
3.2_20	50	-1.05	0.25	0.11	-1.16	0.30	0.13	-1.18	0.48	0.27	-1.46	0.92	0.90
3.2_21	50	-1.05	0.48	0.73	-1.16	0.52	0.97	-1.18	0.48	0.25	-1.46	1.04	0.80
3.2_22	50	-1.05	0.98	1.25	-1.16	1.00	1.65	-1.18	0.89	0.92	-1.46	0.90	0.71
3.2_23	50	-1.05	1.69	2.39	-1.16	1.63	2.61	-1.18	1.33	2.79	-1.46	1.14	1.93

Table 23 (continued): Item statistics of the linear logistic test models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 3.2.p													
3.2_24	50	-1.05	1.49	1.27	-1.16	1.53	1.29	-1.18	1.20	0.94	-1.46	1.10	1.39
Level 4.2.c1													
4.3_1	49	-2.60	0.78	0.73	-2.96	0.72	0.72	-2.88	0.95	1.00	-2.59	1.38	1.40
4.3_2	50	-2.60	1.02	0.89	-2.96	0.89	0.72	-2.88	1.01	0.93	-2.59	1.13	1.01
4.3_5	48	-2.60	0.76	0.59	-2.96	0.73	0.59	-2.88	0.93	0.87	-2.59	1.02	1.55
4.3_7	48	-2.60	0.87	0.86	-2.96	0.86	0.93	-2.88	1.08	1.24	-2.59	1.35	1.45
4.3_9	50	-2.60	1.02	1.05	-2.96	0.99	1.02	-2.88	1.20	1.32	-2.59	0.92	0.86
4.3_10	49	-2.60	0.84	0.76	-2.96	0.81	0.78	-2.88	0.96	0.96	-2.59	0.86	1.41
4.3_13	48	-2.60	0.86	0.66	-2.96	0.81	0.66	-2.88	0.72	0.61	-2.59	0.75	0.58
4.3_15	49	-2.60	1.18	1.25	-2.96	1.10	1.09	-2.88	0.96	1.02	-2.59	1.00	1.35
4.3_17	47	-2.60	1.31	1.46	-2.96	1.26	1.43	-2.88	1.35	1.76	-2.59	1.20	1.35
4.3_18	48	-2.60	0.92	0.74	-2.96	0.86	0.74	-2.88	0.86	0.73	-2.59	0.81	0.63
4.3_21	48	-2.60	0.79	0.70	-2.96	0.78	0.74	-2.88	0.81	0.72	-2.59	0.76	0.58
4.3_23	48	-2.60	1.54	1.85	-2.96	1.45	1.90	-2.88	1.32	1.97	-2.59	1.10	1.18
Level 4.2.c2													
4.3_3	49	-3.84	1.16	1.32	-4.43	1.01	1.17	-4.62	1.02	1.40	-4.31	0.80	0.67
4.3_4	50	-3.84	0.98	0.95	-4.43	0.90	0.99	-4.62	0.99	1.38	-4.31	0.69	1.09
4.3_6	49	-3.84	0.89	0.86	-4.43	0.92	0.91	-4.62	0.72	0.58	-4.31	0.84	0.49
4.3_8	49	-3.84	0.84	0.91	-4.43	0.82	1.00	-4.62	0.88	1.35	-4.31	1.07	0.83
4.3_11	49	-3.84	1.09	1.12	-4.43	1.24	1.33	-4.62	1.17	1.34	-4.31	1.16	0.93
4.3_12	48	-3.84	0.94	0.89	-4.43	0.98	0.96	-4.62	1.00	1.16	-4.31	0.61	1.21
4.3_14	49	-3.84	1.31	1.55	-4.43	1.16	1.42	-4.62	1.08	1.13	-4.31	0.61	0.35
4.3_16	48	-3.84	0.98	1.00	-4.43	0.87	0.88	-4.62	0.99	1.11	-4.31	0.96	0.69
4.3_19	48	-3.84	1.13	1.25	-4.43	1.12	1.33	-4.62	1.11	1.12	-4.31	1.23	1.40
4.3_20	48	-3.84	1.12	1.28	-4.43	1.14	1.46	-4.62	0.69	0.58	-4.31	0.72	0.58
4.3_22	49	-3.84	1.13	1.26	-4.43	1.11	1.48	-4.62	1.22	2.14	-4.31	1.28	2.72
4.3_24	50	-3.84	0.98	0.96	-4.43	0.99	0.90	-4.62	1.05	0.90	-4.31	1.08	0.62

Study 3

Table 24: Item statistics of the Rasch models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.1													
4.1_8	52				2.32	0.88	0.30	2.59	0.68	0.19	1.80	1.15	0.96
4.1_9	52	1.97	1.00	4.77	2.32	1.00	9.20	2.11	0.99	5.00	2.08	1.05	0.82
4.1_14	52	1.97	0.95	0.62	1.58	0.86	0.34	1.75	0.90	0.53	1.94	0.74	0.51
4.1_16	52	1.97	0.97	0.90	1.58	0.92	0.89	1.46	1.04	1.69	1.67	0.79	1.50
4.1_19	52	1.97	0.97	0.90	2.32	0.98	1.57	1.75	0.95	1.85	2.08	1.28	1.67
Level 3.1													
3.1_3	52	0.79	1.06	0.96	1.14	1.02	0.97	1.75	1.42	2.25	2.08	1.09	2.43
3.1_7	52							1.75	0.88	0.47	1.80	0.87	0.64
3.1_8	52	1.97	0.95	0.62	2.32	0.92	0.43	2.11	0.86	0.44	2.23	0.99	8.51
3.1_12	52	1.97	0.94	0.51	1.58	0.89	0.44	2.11	0.83	0.60	2.39	0.99	0.92
3.1_24	52	1.97	0.98	1.25	2.32	0.99	2.23	2.11	1.04	0.47	2.39	0.91	0.74
Level 4.2.p													
4.2_2	52	1.24	0.68	0.22	1.14	0.76	0.29	1.46	0.57	0.22	1.16	1.03	1.26
4.2_7	52	0.47	0.71	0.35	0.33	0.85	0.59	0.15	0.93	0.93	0.08	1.00	0.98
4.2_9	52	0.47	0.86	0.71	0.13	0.80	0.55	0.45	0.75	0.50	0.32	0.85	0.72
4.2_13	51							1.66	0.80	0.42	0.60	0.81	0.63
4.2_22	51	0.73	0.73	0.43	0.76	0.85	0.54	0.70	0.66	0.38	0.60	1.10	1.00
Level 3.2.p													
3.2_1	52	-0.01	0.85	0.75	0.33	0.89	0.95	0.01	0.82	1.05	0.20	1.11	1.05
3.2_4	52	-0.36	0.83	0.73	-0.20	0.81	0.74	0.30	0.89	0.72	-0.17	0.91	0.81
3.2_10	52	-0.66	0.99	1.37	-0.35	1.21	1.64	-0.37	1.24	1.42	-0.84	0.99	0.82
3.2_18	52	0.47	0.81	0.60	0.55	0.83	0.47	0.45	0.75	0.49	0.68	1.00	0.86
3.2_20	52	0.47	0.77	0.49	0.81	0.84	0.60	0.62	0.83	0.76	0.08	0.65	0.49
Level 4.2.c1													
4.3_5	52	-1.43	0.92	0.80	-1.68	1.00	0.89	-1.50	0.73	1.33	-1.15	0.92	1.40
4.3_10	50	-0.58	0.87	0.68	-0.76	0.82	0.64	-1.35	0.85	0.78	-1.30	0.85	1.64
4.3_17	51	-0.67	0.78	0.58	-1.16	0.74	0.61	-1.70	0.63	0.51	-0.56	1.08	1.08
4.3_18	51	-1.15	0.87	0.74	-1.37	0.93	0.83	-1.48	0.79	0.88	-0.84	0.76	0.53
4.3_23	52	-0.66	1.28	1.45	-1.19	1.13	1.28	-1.39	1.08	1.81	-0.99	1.04	0.79
Level 4.2.c2													
4.3_6	52	-2.55	1.00	1.00	-3.35	0.80	0.92	-3.55	1.16	8.43	-4.35	1.11	0.81
4.3_8	51	-2.37	1.08	1.08	-2.91	1.23	1.28	-2.97	1.26	4.68	-3.06	0.94	0.71
4.3_14	51	-2.54	1.21	1.50	-3.01	1.03	1.36	-3.40	1.00	2.19	-3.06	0.85	4.38
4.3_16	51	-3.06	1.15	1.23	-3.12	1.22	1.25	-4.33	1.17	1.55	-4.33	1.10	0.63
4.3_22	52	-2.38	0.99	1.22	-2.43	0.96	1.14	-3.25	0.84	0.72	-3.56	0.55	0.13

Table 25: Item statistics of the linear logistic test models for ACC and all splits of wACC

ACC					wACC: split between 1 and 2			wACC: split between 2 and 3			wACC: split between 3 and 4		
Item	N	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Level 4.1													
4.1_8	52				0.33	0.67	0.24	0.40	0.47	0.16	0.18	1.25	1.08
4.1_9	52	0.70	1.15	5.47	0.33	0.79	6.51	0.40	1.00	4.84	0.18	1.05	0.83
4.1_14	52	0.70	1.10	0.71	0.33	1.27	0.48	0.40	1.16	0.71	0.18	0.78	0.55
4.1_16	52	0.70	1.12	1.06	0.33	1.34	1.31	0.40	1.56	3.05	0.18	0.92	2.01
4.1_19	52	0.70	1.12	1.06	0.33	0.77	1.15	0.40	1.21	2.55	0.18	1.27	1.65
Level 3.1													
3.1_3	52	0.00	1.75	1.77	0.00	1.55	1.58	0.00	1.39	2.14	0.00	1.05	2.06
3.1_7	52							0.00	0.87	0.46	0.00	0.90	0.66
3.1_8	52	0.00	0.62	0.38	0.00	0.55	0.25	0.00	0.68	0.35	0.00	0.93	5.72
3.1_12	52	0.00	0.61	0.33	0.00	0.96	0.47	0.00	0.66	0.45	0.00	0.90	0.88
3.1_24	52	0.00	0.65	0.75	0.00	0.62	1.19	0.00	0.83	0.38	0.00	0.84	0.66
Level 4.2.p													
4.2_2	52	-0.80	0.39	0.17	-1.16	0.49	0.23	-1.05	0.40	0.22	-1.56	1.17	1.14
4.2_7	52	-0.80	0.83	0.40	-1.16	0.97	0.67	-1.05	1.16	1.33	-1.56	0.99	0.91
4.2_9	52	-0.80	0.98	0.80	-1.16	1.06	0.71	-1.05	0.84	0.56	-1.56	0.84	0.71
4.2_13	51							-1.05	0.56	0.32	-1.56	0.81	0.64
4.2_22	51	-0.80	0.67	0.40	-1.16	0.73	0.46	-1.05	0.65	0.38	-1.56	1.09	1.03
Level 3.2.p													
3.2_1	52	-1.50	0.82	0.72	-1.49	0.82	0.84	-1.44	0.90	1.24	-1.74	1.10	1.03
3.2_4	52	-1.50	1.03	0.90	-1.49	1.04	0.95	-1.44	0.88	0.71	-1.74	0.89	0.76
3.2_10	52	-1.50	1.41	2.20	-1.49	1.54	2.47	-1.44	1.50	2.14	-1.74	0.98	0.99
3.2_18	52	-1.50	0.58	0.43	-1.49	0.67	0.40	-1.44	0.70	0.46	-1.74	1.04	0.97
3.2_20	52	-1.50	0.54	0.37	-1.49	0.59	0.42	-1.44	0.74	0.64	-1.74	0.64	0.49
Level 4.2.c1													
4.3_5	52	-2.36	1.17	1.00	-2.92	1.12	0.99	-3.19	0.73	1.25	-2.87	0.87	1.18
4.3_10	50	-2.36	0.75	0.60	-2.92	0.73	0.61	-3.19	0.84	0.78	-2.87	0.80	1.22
4.3_17	51	-2.36	0.69	0.54	-2.92	0.72	0.60	-3.19	0.64	0.52	-2.87	1.19	1.31
4.3_18	51	-2.36	0.98	0.78	-2.92	0.97	0.84	-3.19	0.78	0.86	-2.87	0.78	0.55
4.3_23	52	-2.36	1.15	1.23	-2.92	1.11	1.23	-3.19	1.07	1.78	-2.87	1.01	0.76
Level 4.2.c2													
4.3_6	52	-4.00	1.00	1.00	-4.61	0.76	0.80	-5.14	1.12	6.96	-5.43	0.75	0.41
4.3_8	51	-4.00	1.08	1.11	-4.61	1.22	1.28	-5.14	1.43	6.59	-5.43	1.26	1.07
4.3_14	51	-4.00	1.20	1.48	-4.61	1.01	1.29	-5.14	1.01	2.18	-5.43	1.17	6.49
4.3_16	51	-4.00	1.14	1.17	-4.61	1.18	1.19	-5.14	0.94	0.93	-5.43	0.75	0.34
4.3_22	52	-4.00	1.00	1.28	-4.61	1.06	1.47	-5.14	0.90	0.80	-5.43	0.56	0.13

Results of Study 4

Location of fixations

Table 26: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 4.1

Item	Item sample	N	L0	L.o	L.n	L	R0	R.o.nr	R.o.r	R.n.nr	R.n.r
4.1_8	all items	48	.12 (.08,.16)	.16 (.12,.20)	.05 (.03,.07)	.34 (.29,.39)	.18 (.14,.22)	.08 (.05,.11)	.16 (.13,.19)	.06 (.03,.09)	.19 (.15,.23)
	incorrect	2	.18 (-.17,.53)	.03 (-.03,.09)	.12 (-.11,.35)	.32 (-.31,.95)	.12 (-.11,.35)	.20 (-.19,.59)	.13 (-.01,.27)	.20 (-.19,.59)	.03 (-.03,.09)
	correct	46	.12 (.08,.16)	.17 (.13,.21)	.05 (.03,.07)	.34 (.30,.38)	.18 (.14,.22)	.07 (.05,.09)	.16 (.13,.19)	.06 (.03,.09)	.20 (.16,.24)
4.1_9	all items	48	.21 (.18,.24)	.10 (.07,.13)	.04 (.02,.06)	.35 (.31,.39)	.22 (.17,.27)	.07 (.04,.10)	.17 (.13,.21)	.06 (.03,.09)	.13 (.10,.16)
	incorrect	5	.19 (.11,.27)	.03 (-.01,.07)	.01 (-.02,.04)	.24 (.13,.35)	.29 (.12,.46)	.04 (-.01,.09)	.19 (.06,.32)	.04 (-.01,.09)	.20 (.19,.21)
	correct	43	.22 (.18,.26)	.10 (.07,.13)	.05 (.03,.07)	.37 (.32,.42)	.21 (.16,.26)	.07 (.04,.10)	.16 (.12,.20)	.06 (.03,.09)	.12 (.09,.15)
4.1_14	all items	46	.02 (.01,.03)	.14 (.11,.17)	.16 (.12,.20)	.32 (.28,.36)	.05 (.03,.07)	.24 (.20,.28)	.10 (.07,.13)	.14 (.10,.18)	.15 (.12,.18)
	incorrect	2	.00 (.00,.00)	.00 (.00,.00)	.17 (.16,.18)	.17 (.16,.18)	.05 (-.04,.14)	.17 (.16,.18)	.14 (-.13,.41)	.33 (-.32,.98)	.14 (-.13,.41)
	correct	44	.03 (.01,.05)	.14 (.11,.17)	.16 (.12,.20)	.33 (.29,.37)	.05 (.03,.07)	.24 (.20,.28)	.09 (.06,.12)	.14 (.11,.17)	.15 (.12,.18)
4.1_16	all items	46	.03 (.01,.05)	.10 (.06,.14)	.18 (.14,.22)	.31 (.27,.35)	.18 (.14,.22)	.04 (.02,.06)	.19 (.15,.23)	.07 (.05,.09)	.20 (.16,.24)
	incorrect	3	.06 (.00,.12)	.17 (-.10,.44)	.03 (-.01,.07)	.26 (-.05,.57)	.18 (.00,.36)	.00 (-.01,.01)	.18 (.02,.34)	.09 (-.03,.21)	.29 (.11,.47)
	correct	43	.03 (.01,.05)	.10 (.07,.13)	.19 (.15,.23)	.32 (.28,.36)	.18 (.14,.22)	.05 (.03,.07)	.20 (.16,.24)	.06 (.03,.09)	.20 (.16,.24)
4.1_19	all items	48	.04 (.02,.06)	.05 (.03,.07)	.17 (.13,.21)	.26 (.22,.30)	.05 (.03,.07)	.21 (.17,.25)	.16 (.12,.20)	.16 (.13,.19)	.16 (.13,.19)
	incorrect	1	.00 (.00,.00)	.07 (.00,.14)	.13 (.00,.26)	.20 (.00,.40)	.00 (.00,.00)	.00 (.00,.00)	.33 (.00,.66)	.20 (.00,.40)	.27 (.00,.54)
	correct	47	.04 (.02,.06)	.05 (.03,.07)	.17 (.13,.21)	.26 (.22,.30)	.05 (.03,.07)	.21 (.17,.25)	.16 (.12,.20)	.16 (.13,.19)	.16 (.13,.19)

Table 27: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 3.1

Item	Item sample	N	L0	L.o	L.n	L	R0	R.o.nr	R.o.r	R.n.nr	R.n.r
3.1_3	all items	46	.06	.18	.07	.32	.09	.12	.17	.16	.14
			(.03,.09)	(.15,.21)	(.04,.10)	(.27,.37)	(.07,.11)	(.08,.16)	(.13,.21)	(.12,.20)	(.10,.18)
	incorrect	2	.00	.20	.09	.29	.03	.35	.20	.00	.12
3.1_7	correct	44	.07	.18	.07	.32	.09	.11	.17	.16	.14
			(.04,.10)	(.15,.21)	(.04,.10)	(.27,.37)	(.06,.12)	(.08,.14)	(.13,.21)	(.12,.20)	(.10,.18)
	incorrect	4	.00	.28	.00	.28	.03	.21	.13	.19	.16
3.1_8	all items	45	.02	.11	.18	.31	.07	.16	.17	.08	.20
			(.01,.03)	(.07,.15)	(.14,.22)	(.27,.35)	(.04,.10)	(.13,.19)	(.13,.21)	(.05,.11)	(.17,.23)
	incorrect	2	.00	.00	.14	.14	.12	.14	.17	.23	.21
3.1_12	correct	43	.02	.12	.18	.32	.07	.16	.17	.07	.20
			(.01,.03)	(.08,.16)	(.14,.22)	(.28,.36)	(.04,.10)	(.13,.19)	(.13,.21)	(.05,.09)	(.16,.24)
	incorrect	2	.10	.17	.07	.33	.03	.13	.03	.20	.27
3.1_24	all items	46	.06	.05	.20	.30	.05	.17	.16	.12	.20
			(.04,.08)	(.03,.07)	(.16,.24)	(.26,.34)	(.03,.07)	(.13,.21)	(.11,.21)	(.08,.16)	(.17,.23)
	incorrect	7	.03	.05	.17	.25	.06	.13	.19	.08	.28
	correct	39	.06	.04	.20	.31	.05	.18	.16	.12	.19
			(.04,.08)	(.02,.06)	(.16,.24)	(.26,.36)	(.03,.07)	(.13,.23)	(.11,.21)	(.08,.16)	(.15,.23)

Table 28: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 4.2.p

Item	Itemsample	N	L0	L.o1	L.o2	L.o	L	R0	R.o.nr	R.o.o	R.o.n
4.2_2	all items	51	.02 (.01,.03)	.10 (.07,.13)	.15 (.11,.19)	.25 (.20,.30)	.27 (.22,.32)	.10 (.07,.13)	.33 (.29,.37)	.12 (.09,.15)	.19 (.16,.22)
	incorrect	9	.01 (-.01,.03)	.08 (.02,.14)	.09 (-.02,.20)	.18 (.07,.29)	.19 (.08,.30)	.13 (.03,.23)	.38 (.30,.46)	.10 (.03,.17)	.20 (.12,.28)
	correct	42	.02 (.01,.03)	.11 (.08,.14)	.16 (.11,.21)	.27 (.22,.32)	.29 (.24,.34)	.09 (.06,.12)	.32 (.27,.37)	.12 (.09,.15)	.19 (.16,.22)
4.2_7	all items	52	.08 (.06,.10)	.13 (.10,.16)	.06 (.04,.08)	.19 (.15,.23)	.27 (.23,.31)	.08 (.06,.10)	.10 (.07,.13)	.31 (.27,.35)	.24 (.21,.27)
	incorrect	6	.06 (-.01,.13)	.10 (.01,.19)	.02 (.00,.04)	.12 (.02,.22)	.18 (.02,.34)	.06 (.00,.12)	.21 (.12,.30)	.25 (.17,.33)	.30 (.22,.38)
	correct	46	.08 (.05,.11)	.14 (.11,.17)	.06 (.03,.09)	.20 (.16,.24)	.28 (.24,.32)	.08 (.05,.11)	.08 (.05,.11)	.32 (.28,.36)	.24 (.20,.28)
4.2_9	all items	50	.06 (.04,.08)	.07 (.05,.09)	.14 (.10,.18)	.21 (.17,.25)	.27 (.22,.32)	.03 (.02,.04)	.19 (.16,.22)	.26 (.22,.30)	.25 (.21,.29)
	incorrect	10	.11 (.04,.18)	.03 (.01,.05)	.13 (.04,.22)	.16 (.06,.26)	.27 (.15,.39)	.04 (.01,.07)	.23 (.15,.31)	.23 (.14,.32)	.24 (.12,.36)
	correct	40	.05 (.03,.07)	.08 (.05,.11)	.14 (.10,.18)	.22 (.17,.27)	.27 (.22,.32)	.02 (.01,.03)	.19 (.16,.22)	.27 (.22,.32)	.25 (.21,.29)
4.2_13	all items	51	.10 (.08,.12)	.15 (.11,.19)	.06 (.04,.08)	.21 (.16,.26)	.30 (.26,.34)	.06 (.04,.08)	.27 (.24,.30)	.18 (.16,.20)	.19 (.16,.22)
	incorrect	4	.18 (.03,.33)	.06 (-.02,.14)	.08 (-.03,.19)	.15 (.02,.28)	.33 (.10,.56)	.05 (-.01,.11)	.26 (.15,.37)	.13 (.04,.22)	.23 (.12,.34)
	correct	47	.09 (.07,.11)	.16 (.12,.20)	.06 (.04,.08)	.21 (.16,.26)	.30 (.26,.34)	.06 (.04,.08)	.27 (.24,.30)	.19 (.17,.21)	.18 (.15,.21)
4.2_22	all items	49	.07 (.05,.09)	.07 (.05,.09)	.16 (.13,.19)	.24 (.20,.28)	.31 (.27,.35)	.08 (.05,.11)	.34 (.30,.38)	.13 (.10,.16)	.14 (.12,.16)
	incorrect	6	.04 (-.01,.09)	.00 (.00,.00)	.18 (.11,.25)	.18 (.11,.25)	.22 (.16,.28)	.07 (-.01,.15)	.37 (.25,.49)	.19 (.10,.28)	.16 (.06,.26)
	correct	43	.08 (.06,.10)	.08 (.05,.11)	.16 (.12,.20)	.24 (.19,.29)	.32 (.27,.37)	.09 (.06,.12)	.33 (.29,.37)	.12 (.09,.15)	.14 (.12,.16)

Table 29: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 3.2.p

Item	Item sample	N	L0	L0.1	L0.2	L.o	L	R0	R.o.nr	R.o.o	R.o.n
3.2_1	all items	51	.10 (.08,.12)	.06 (.04,.08)	.16 (.12,.20)	.22 (.18,.26)	.32 (.28,.36)	.06 (.04,.08)	.04 (.02,.06)	.30 (.26,.34)	.28 (.25,.31)
	incorrect	10	.10	.09	.11	.20	.29	.07	.03	.30	.30
	correct	41	.10 (.07,.13)	.05 (.01,.17)	.17 (.02,.20)	.22 (.10,.30)	.32 (.18,.40)	.06 (.02,.12)	.05 (.00,.06)	.30 (.22,.38)	.28 (.25,.35)
3.2_4	all items	50	.06 (.04,.12)	.07 (.03,.07)	.19 (.13,.21)	.26 (.18,.26)	.32 (.28,.36)	.05 (.04,.08)	.29 (.03,.07)	.19 (.25,.35)	.14 (.24,.32)
	incorrect	9	.08	.05	.11	.16	.25	.07	.35	.22	.12
	correct	41	.06 (.04,.08)	.08 (.06,.10)	.20 (.16,.24)	.28 (.23,.33)	.34 (.30,.38)	.05 (.03,.07)	.28 (.25,.31)	.18 (.15,.21)	.14 (.11,.17)
3.2_10	all items	52	.06 (.04,.08)	.16 (.13,.19)	.06 (.04,.08)	.22 (.18,.26)	.28 (.24,.32)	.07 (.05,.09)	.07 (.05,.09)	.30 (.27,.33)	.28 (.25,.31)
	incorrect	11	.05	.13	.03	.17	.22	.09	.08	.32	.30
	correct	41	.06 (.04,.08)	.17 (.14,.20)	.07 (.05,.09)	.24 (.19,.29)	.30 (.25,.35)	.07 (.05,.09)	.06 (.03,.09)	.30 (.26,.34)	.27 (.23,.31)
3.2_18	all items	52	.08 (.06,.10)	.15 (.11,.19)	.06 (.04,.08)	.22 (.18,.26)	.30 (.26,.34)	.05 (.03,.07)	.04 (.02,.06)	.32 (.28,.36)	.29 (.25,.33)
	incorrect	6	.06	.17	.01	.18	.24	.12	.03	.31	.31
	correct	46	.08 (.06,.10)	.15 (.11,.19)	.07 (.05,.09)	.22 (.18,.26)	.30 (.25,.35)	.05 (.03,.07)	.05 (.03,.07)	.32 (.28,.36)	.29 (.25,.33)
3.2_20	all items	50	.08 (.06,.10)	.17 (.14,.20)	.07 (.05,.09)	.23 (.19,.27)	.31 (.27,.35)	.07 (.05,.09)	.25 (.22,.28)	.17 (.14,.20)	.20 (.18,.22)
	incorrect	11	.10	.16	.03	.19	.29	.07	.24	.18	.23
	correct	39	.08 (.04,.16)	.17 (.07,.25)	.08 (.00,.06)	.24 (.09,.29)	.32 (.20,.38)	.07 (.02,.12)	.25 (.14,.34)	.16 (.12,.24)	.20 (.16,.30)
			.08 (.06,.10)	.17 (.13,.21)	.08 (.06,.10)	.24 (.20,.28)	.32 (.28,.36)	.07 (.04,.10)	.25 (.22,.28)	.16 (.13,.19)	.20 (.18,.22)

Table 30: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 4.2.c1

Item	Item sample N	L.p	L.c	L1	L	R.p.nr	R.p.r	R.c.nr1	R.c.nr2	R.c.r1	R.c.r2	R1.nr	R1.r1.n/ R1.r1.o	R.ele- ment1.n	R.ele- ment1.o
4.3_5	all items	51	.06	.07	.03	.16	.23	.08	.17	.32	.15	.08	.06	.06	.15
			(.04,.08)	(.05,.09)	(.02,.04)	(.12,.20)	(.19,.27)	(.06,.10)	(.14,.20)	(.28,.36)	(.12,.18)	(.06,.10)	(.04,.08)	(.04,.08)	(.12,.18)
	incorrect	14	.10	.08	.03	.21	.29	.07	.11	.24	.13	.05	.08	.06	.13
	correct	37	.04	.07	.03	.14	.20	.08	.19	.35	.16	.10	.08	.05	.16
4.3_10	all items	51	.01	.09	.06	.16	.12	.11	.20	.31	.11	.16	.06	.07	.11
			(.02,.06)	(.04,.10)	(.01,.05)	(.10,.18)	(.16,.24)	(.05,.11)	(.16,.22)	(.31,.39)	(.13,.19)	(.07,.13)	(.06,.10)	(.03,.07)	(.13,.19)
	incorrect	12	.01	.13	.10	.24	.16	.14	.15	.19	.04	.16	.06	.04	.04
	correct	39	.02	.07	.04	.13	.11	.10	.21	.34	.13	.17	.06	.08	.13
4.3_17	all items	52	.05	.04	.04	.13	.10	.15	.12	.20	.09	.18	.14	.09	.09
			(.01,.03)	(.04,.10)	(.02,.06)	(.09,.17)	(.08,.14)	(.07,.13)	(.17,.25)	(.30,.38)	(.10,.16)	(.13,.21)	(.04,.08)	(.06,.10)	(.10,.16)
	incorrect	19	.04	.03	.03	.10	.10	.14	.12	.21	.09	.15	.21	.09	.09
	correct	33	.05	.05	.05	.15	.10	.16	.12	.20	.08	.19	.10	.09	.09
4.3_18	all items	52	.05	.08	.04	.18	.09	.10	.31	.43	.12	.07	.06	.06	.12
			(.03,.07)	(.02,.08)	(.03,.07)	(.10,.20)	(.07,.13)	(.12,.20)	(.09,.15)	(.16,.24)	(.05,.11)	(.15,.23)	(.06,.14)	(.06,.12)	(.06,.12)
	incorrect	15	.07	.09	.03	.19	.12	.06	.29	.43	.14	.04	.09	.08	.14
	correct	38	.04	.07	.05	.16	.08	.12	.33	.44	.11	.08	.05	.05	.11
4.3_23	all items	52	.03	.10	.02	.15	.05	.17	.21	.25	.03	.28	.09	.02	.02
			(.02,.06)	(.04,.10)	(.03,.07)	(.11,.21)	(.06,.10)	(.10,.14)	(.28,.38)	(.40,.48)	(.07,.15)	(.04,.12)	(.04,.08)	(.03,.07)	(.07,.15)
	incorrect	12	.05	.13	.01	.19	.07	.17	.19	.23	.03	.23	.09	.02	.03
	correct	40	.02	.09	.02	.14	.05	.17	.22	.25	.04	.29	.09	.02	.02
			(.01,.03)	(.06,.12)	(.01,.03)	(.10,.18)	(.03,.07)	(.14,.20)	(.19,.25)	(.21,.29)	(.03,.05)	(.24,.34)	(.07,.11)	(.01,.03)	(.03,.05)
			(.01,.09)	(.06,.20)	(-.01,.03)	(.11,.27)	(.01,.13)	(.11,.23)	(.13,.25)	(.16,.30)	(-.01,.07)	(.16,.30)	(.04,.14)	(.00,.04)	(.00,.04)
			(.01,.03)	(.06,.12)	(.01,.03)	(.10,.18)	(.03,.07)	(.14,.20)	(.19,.25)	(.21,.29)	(.03,.05)	(.24,.34)	(.07,.11)	(.01,.03)	(.03,.05)

Table 31: Mean (95%-confidence interval) of the relative number of fixations of AOIs of Level 4.2.c2

Item	Item sample	N	L.p	L.c	L1	L	R.p.nr	R.p.r	R.c.nr1	R.c.nr2	R.c.r1	R.c.r2	R1.nr	R1.r1.n/ R1.r1.o	Rele- ment1.n	Rele- ment1.o
4.3_6	<i>all items</i>	52	.05 (.03,.07)	.04 (.02,.06)	.04 (.02,.06)	.14 (.10,.18)	.17 (.14,.20)	.14 (.10,.18)	.11 (.09,.13)	.20 (.17,.23)	.09 (.07,.11)	.14 (.11,.17)	.13 (.11,.15)	.09 (.06,.12)	.09 (.06,.12)	.09 (.07,.11)
	<i>incorrect</i>	31	.07 (.03,.11)	.05 (.03,.07)	.03 (.01,.05)	.15 (.10,.20)	.20 (.16,.24)	.12 (.08,.16)	.10 (.07,.13)	.18 (.14,.22)	.09 (.06,.12)	.13 (.09,.17)	.12 (.09,.15)	.09 (.06,.12)	.09 (.06,.12)	.09 (.06,.12)
	<i>correct</i>	21	.02 (.00,.04)	.04 (.02,.06)	.05 (.01,.09)	.11 (.05,.17)	.12 (.08,.16)	.16 (.10,.22)	.13 (.10,.16)	.23 (.18,.28)	.10 (.06,.14)	.15 (.11,.19)	.14 (.10,.18)	.09 (.04,.14)	.09 (.04,.14)	.10 (.06,.14)
4.3_8	<i>all items</i>	51	.07 (.05,.09)	.05 (.02,.08)	.06 (.04,.08)	.18 (.13,.23)	.23 (.19,.27)	.07 (.05,.09)	.22 (.19,.25)	.28 (.25,.31)	.06 (.04,.08)	.05 (.03,.07)	.16 (.14,.18)	.04 (.03,.05)	.06 (.04,.08)	.04 (.03,.05)
	<i>incorrect</i>	30	.08 (.05,.11)	.06 (.02,.10)	.06 (.04,.08)	.20 (.15,.25)	.22 (.17,.27)	.07 (.04,.10)	.21 (.17,.25)	.27 (.22,.32)	.07 (.04,.10)	.04 (.03,.05)	.16 (.13,.19)	.04 (.02,.06)	.07 (.04,.10)	.04 (.02,.06)
	<i>correct</i>	21	.05 (.02,.08)	.05 (.01,.09)	.05 (.01,.09)	.15 (.07,.23)	.23 (.17,.29)	.07 (.03,.11)	.23 (.19,.27)	.29 (.25,.33)	.06 (.04,.08)	.06 (.03,.09)	.17 (.13,.21)	.03 (.01,.05)	.06 (.04,.08)	.03 (.01,.05)
4.3_14	<i>all items</i>	51	.03 (.01,.05)	.06 (.04,.08)	.04 (.03,.05)	.13 (.10,.16)	.19 (.16,.22)	.07 (.05,.09)	.15 (.12,.18)	.25 (.21,.29)	.10 (.07,.13)	.10 (.07,.13)	.19 (.16,.22)	.08 (.06,.10)	.08 (.06,.10)	.10 (.07,.13)
	<i>incorrect</i>	32	.04 (.02,.06)	.07 (.04,.10)	.04 (.02,.06)	.14 (.10,.18)	.19 (.17,.21)	.06 (.04,.08)	.15 (.12,.18)	.26 (.22,.30)	.11 (.08,.14)	.08 (.05,.11)	.18 (.14,.22)	.08 (.05,.11)	.08 (.05,.11)	.11 (.08,.14)
	<i>correct</i>	19	.03 (.01,.05)	.04 (.02,.06)	.04 (.01,.07)	.11 (.05,.17)	.19 (.13,.25)	.08 (.05,.11)	.15 (.11,.19)	.24 (.16,.32)	.09 (.04,.14)	.12 (.07,.17)	.19 (.14,.24)	.08 (.05,.11)	.08 (.05,.11)	.09 (.04,.14)
4.3_16	<i>all items</i>	52	.05 (.03,.07)	.06 (.04,.08)	.05 (.03,.07)	.16 (.12,.20)	.16 (.13,.19)	.08 (.05,.11)	.23 (.19,.27)	.31 (.27,.35)	.08 (.05,.11)	.09 (.06,.12)	.13 (.10,.16)	.07 (.05,.09)	.08 (.05,.11)	.07 (.05,.09)
	<i>incorrect</i>	32	.07 (.04,.10)	.07 (.05,.09)	.05 (.03,.07)	.19 (.14,.24)	.16 (.12,.20)	.09 (.05,.13)	.22 (.18,.26)	.28 (.23,.33)	.06 (.03,.09)	.10 (.06,.14)	.12 (.08,.16)	.06 (.03,.09)	.06 (.03,.09)	.06 (.03,.09)
	<i>correct</i>	20	.02 (.00,.04)	.06 (.03,.09)	.05 (.02,.08)	.13 (.07,.19)	.15 (.11,.19)	.07 (.05,.09)	.24 (.17,.31)	.35 (.28,.42)	.11 (.05,.17)	.07 (.03,.11)	.14 (.10,.18)	.10 (.05,.15)	.11 (.05,.17)	.10 (.05,.15)
4.3_22	<i>all items</i>	51	.06 (.04,.08)	.04 (.02,.06)	.04 (.02,.06)	.13 (.09,.17)	.16 (.13,.19)	.14 (.11,.17)	.14 (.11,.17)	.21 (.18,.24)	.07 (.05,.09)	.09 (.06,.12)	.19 (.16,.22)	.07 (.05,.09)	.07 (.05,.09)	.07 (.05,.09)
	<i>incorrect</i>	29	.06 (.03,.09)	.04 (.01,.07)	.04 (.02,.06)	.14 (.09,.19)	.15 (.11,.19)	.15 (.11,.19)	.15 (.11,.19)	.20 (.16,.24)	.05 (.03,.07)	.10 (.06,.14)	.19 (.15,.23)	.07 (.04,.10)	.07 (.04,.10)	.05 (.03,.07)
	<i>correct</i>	22	.05 (.03,.09)	.03 (.01,.07)	.04 (.02,.06)	.12 (.09,.19)	.17 (.11,.19)	.14 (.11,.19)	.13 (.11,.19)	.22 (.16,.24)	.09 (.03,.07)	.08 (.06,.14)	.19 (.15,.23)	.08 (.04,.10)	.08 (.04,.10)	.09 (.03,.07)

Entropy

Table 32: Mean (95%-confidence interval) of the entropy values of Level 4.1 and 3.1

Level 4.1				Level 3.1		
Item sample	Item	N	M(SD)	Item	N	M(SD)
all items	4.1_8	47	0.45 (0.40,0.50)	3.1_3	46	0.44 (0.39,0.49)
incorrect		1	0.66 -		2	0.25 (0.25,0.25)
correct		46	0.45 (0.40,0.50)		44	0.45 (0.40,0.50)
all items	4.1_9	48	0.47 (0.42,0.52)	3.1_7	47	0.49 (0.45,0.53)
incorrect		5	0.36 (0.20,0.52)		4	0.47 (0.20,0.74)
correct		43	0.49 (0.44,0.54)		43	0.49 (0.45,0.53)
all items	4.1_14	46	0.46 (0.41,0.51)	3.1_8	44	0.45 (0.40,0.50)
incorrect		2	0.20 (0.18,0.22)		2	0.11 (0.04,0.18)
correct		44	0.48 (0.43,0.53)		42	0.46 (0.41,0.51)
all items	4.1_16	45	0.43 (0.38,0.48)	3.1_12	45	0.45 (0.39,0.51)
incorrect		3	0.30 (-0.07,0.67)		2	0.61 (0.55,0.67)
correct		42	0.44 (0.39,0.49)		43	0.44 (0.38,0.50)
all items	4.1_19	44	0.43 (0.38,0.48)	3.1_24	45	0.43 (0.38,0.48)
incorrect		1	0.23 -		7	0.40 (0.27,0.53)
correct		43	0.43 (0.38,.48)		38	0.44 (0.38,0.50)

Table 33: Mean (95%-confidence interval) of the entropy values of Level 4.2.p and 3.2.p

Level 4.2.p				Level 3.2.p		
Item sample	Item	N	M(SD)	Item	N	M(SD)
all items	4.2_2	50	0.37 (0.31,0.43)	3.2_1	51	0.44 (0.39,0.49)
incorrect		8	0.32 (0.20,0.44)		10	0.40 (0.28,0.52)
correct		42	0.38 (0.32,0.44)		41	0.45 (0.40,0.50)
all items	4.2_7	49	0.41 (0.35,0.47)	3.2_4	50	0.48 (0.44,0.52)
incorrect		4	0.33 (0.13,0.53)		9	0.35 (0.25,0.45)
correct		45	0.42 (0.36,0.48)		41	0.50 (0.46,0.54)
all items	4.2_9	47	0.40 (0.34,0.46)	3.2_10	51	0.42 (0.37,0.47)
incorrect		8	0.45 (0.32,0.58)		10	0.43 (0.37,0.49)
correct		39	0.39 (0.33,0.45)		41	0.42 (0.36,0.48)
all items	4.2_13	51	0.40 (0.35,0.45)	3.2_18	51	0.42 (0.36,0.48)
incorrect		4	0.35 (0.13,0.57)		6	0.36 (0.25,0.47)
correct		47	0.40 (0.35,0.45)		45	0.43 (0.37,0.49)
all items	4.2_22	49	0.43 (0.38,0.48)	3.2_20	50	0.46 (0.42,0.50)
incorrect		6	0.37 (0.20,0.54)		11	0.42 (0.33,0.51)
correct		43	0.44 (0.38,0.50)		39	0.47 (0.42,0.52)

Table 34: Mean (95%-confidence interval) of the entropy values of Level 4.2.c1 and 4.2.c2

Level 4.2.c1					Level 4.2.c2			
Item sample	Item	N	M(SD)		Item	N	M(SD)	
all items	4.3_5	38	0.32	(0.26,0.38)	4.3_6	41	0.22	(0.16,0.28)
incorrect		10	0.41	(0.27,0.55)		23	0.25	(0.17,0.33)
correct		28	0.29	(0.22,0.36)		18	0.19	(0.10,0.28)
all items	4.3_10	37	0.31	(0.25,0.37)	4.3_8	41	0.30	(0.24,0.36)
incorrect		9	0.39	(0.30,0.48)		27	0.31	(0.24,0.38)
correct		28	0.28	(0.22,0.34)		14	0.26	(0.15,0.37)
all items	4.3_17	36	0.29	(0.23,0.35)	4.3_14	42	0.26	(0.20,0.32)
incorrect		12	0.24	(0.14,0.34)		27	0.30	(0.23,0.37)
correct		24	0.31	(0.24,0.38)		15	0.21	(0.11,0.31)
all items	4.3_18	37	0.32	(0.26,0.38)	4.3_16	45	0.28	(0.22,0.34)
incorrect		11	0.31	(0.19,0.43)		29	0.30	(0.22,0.38)
correct		26	0.33	(0.27,0.39)		16	0.25	(0.16,0.34)
all items	4.3_23	37	0.29	(0.23,0.35)	4.3_22	42	0.25	(0.19,0.31)
incorrect		9	0.31	(0.17,0.45)		23	0.27	(0.19,0.35)
correct		28	0.29	(0.23,0.35)		19	0.23	(0.15,0.31)

Bi-grams

‘Comparison I’ means comparison between the part of the rotated element with the new color and the corresponding not rotated element on the other side. ‘Comparison II’ means the comparison between the part of the rotated element with the old color and the corresponding not-rotated elements on the other side.

Table 35: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 4.1

Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.1_8a	all items	48	.27	(.14,.40)	.21	(.09,.33)	.25	(.13,.37)
	incorrect	2	.00	(.00,.00)	.50	(-.48,1.48)	.00	(.00,.00)
	correct	46	.28	(.15,.41)	.20	(.08,.32)	.26	(.13,.39)
4.1_9a	all items	48	.06	(-.01,.13)	.19	(.08,.30)	.23	(.11,.35)
	incorrect	5	.00	(.00,.00)	.00	(.00,.00)	.60	(.12,1.08)
	correct	43	.07	(-.01,.15)	.21	(.09,.33)	.19	(.07,.31)
4.1_14a	all items	46	.52	(.37,.67)	.20	(.08,.32)	.24	(.12,.36)
	incorrect	2	.50	(-.48,1.48)	.00	(.00,.00)	.50	(-.48,1.48)
	correct	44	.52	(.37,.67)	.20	(.08,.32)	.23	(.10,.36)
4.1_16a	all items	46	.11	(.02,.20)	.26	(.13,.39)	.39	(.25,.53)
	incorrect	3	.00	(.00,.00)	.33	(-.32,.98)	.33	(-.32,.98)
	correct	43	.12	(.02,.22)	.26	(.13,.39)	.40	(.25,.55)
4.1_19a	all items	48	.52	(.38,.66)	.33	(.20,.46)	.35	(.21,.49)
	incorrect	1	.00	-	1.00	-	1.00	-
	correct	47	.53	(.39,.67)	.32	(.19,.45)	.34	(.20,.48)

Table 36: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 3.1

Item	Item sample	N	Comparison I		Comparison II		Rotation	
3.1_3a	<i>all items</i>	46	.33	(.19,.47)	.33	(.19,.47)	.26	(.13,.39)
	<i>incorrect</i>	2	.50	(-.48,1.48)	.00	(.00,.00)	.00	(.00,.00)
	<i>correct</i>	44	.32	(.18,.46)	.34	(.20,.48)	.27	(.14,.40)
3.1_7a	<i>all items</i>	47	.36	(.22,.50)	.43	(.29,.57)	.17	(.06,.28)
	<i>incorrect</i>	4	1.00	(1.00,1.00)	.50	(-.07,1.07)	.25	(-.24,.74)
	<i>correct</i>	43	.30	(.16,.44)	.42	(.27,.57)	.16	(.05,.27)
3.1_8a	<i>all items</i>	45	.58	(.43,.73)	.27	(.14,.40)	.53	(.38,.68)
	<i>incorrect</i>	2	.50	(-.48,1.48)	.50	(-.48,1.48)	1.00	(1.00,1.00)
	<i>correct</i>	43	.58	(.43,.73)	.26	(.13,.39)	.51	(.36,.66)
3.1_12a	<i>all items</i>	46	.37	(.23,.51)	.24	(.12,.36)	.20	(.08,.32)
	<i>incorrect</i>	2	.00	(.00,.00)	.00	(.00,.00)	.00	(.00,.00)
	<i>correct</i>	44	.39	(.24,.54)	.25	(.12,.38)	.20	(.08,.32)
3.1_24a	<i>all items</i>	46	.46	(.31,.61)	.35	(.21,.49)	.52	(.37,.67)
	<i>incorrect</i>	7	.43	(.03,.83)	.57	(.17,.97)	.86	(.58,1.14)
	<i>correct</i>	39	.46	(.30,.62)	.31	(.16,.46)	.46	(.30,.62)

Table 37: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 4.2.p

Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.2_7a	<i>all items</i>	52	.29	(.17,.41)	.31	(.18,.44)	.65	(.52,.78)
	<i>incorrect</i>	6	.50	(.06,.94)	.50	(.06,.94)	.50	(.06,.94)
	<i>correct</i>	46	.26	(.13,.39)	.28	(.15,.41)	.67	(.53,.81)
4.2_9a	<i>all items</i>	50	.48	(.34,.62)	.38	(.24,.52)	.54	(.40,.68)
	<i>incorrect</i>	10	.30	(.00,.60)	.20	(-.06,.46)	.50	(.17,.83)
	<i>correct</i>	40	.52	(.36,.68)	.42	(.26,.58)	.55	(.39,.71)
4.2_13a	<i>all items</i>	51	.55	(.41,.69)	.53	(.39,.67)	.27	(.15,.39)
	<i>incorrect</i>	4	.75	(.26,1.24)	.50	(-.07,1.07)	.00	(.00,.00)
	<i>correct</i>	47	.53	(.39,.67)	.53	(.39,.67)	.30	(.17,.43)
4.2_22a	<i>all items</i>	49	.37	(.23,.51)	.41	(.27,.55)	.27	(.15,.39)
	<i>incorrect</i>	6	.17	(-.16,.50)	.33	(-.08,.74)	.33	(-.08,.74)
	<i>correct</i>	43	.40	(.25,.55)	.42	(.27,.57)	.26	(.13,.39)
4.2_2a	<i>all items</i>	51	.63	(.50,.76)	.24	(.12,.36)	.18	(.07,.29)
	<i>incorrect</i>	9	.78	(.49,1.07)	.33	(.00,.66)	.22	(-.07,.51)
	<i>correct</i>	42	.60	(.45,.75)	.21	(.08,.34)	.17	(.06,.28)

Table 38: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 3.2.p

Item	Item sample	N	Comparison I		Comparison II		Rotation	
3.2_1a	<i>all items</i>	51	.16	(.06,.26)	.14	(.04,.24)	.65	(.52,.78)
	<i>incorrect</i>	10	.20	(-.06,.46)	.20	(-.06,.46)	.40	(.08,.72)
	<i>correct</i>	41	.15	(.04,.26)	.12	(.02,.22)	.71	(.57,.85)
3.2_4a	<i>all items</i>	50	.22	(.10,.34)	.46	(.32,.60)	.28	(.15,.41)
	<i>incorrect</i>	9	.33	(.00,.66)	.56	(.22,.90)	.44	(.10,.78)
	<i>correct</i>	41	.20	(.08,.32)	.44	(.29,.59)	.24	(.11,.37)
3.2_10a	<i>all items</i>	52	.37	(.24,.50)	.25	(.13,.37)	.67	(.54,.80)
	<i>incorrect</i>	11	.27	(-.01,.55)	.18	(-.06,.42)	.82	(.58,1.06)
	<i>correct</i>	41	.39	(.24,.54)	.27	(.13,.41)	.63	(.48,.78)
3.2_18a	<i>all items</i>	52	.04	(-.01,.09)	.17	(.07,.27)	.62	(.49,.75)
	<i>incorrect</i>	6	.00	(.00,.00)	.17	(-.16,.50)	.50	(.06,.94)
	<i>correct</i>	46	.04	(-.02,.10)	.17	(.06,.28)	.63	(.49,.77)
3.2_20a	<i>all items</i>	50	.48	(.34,.62)	.48	(.34,.62)	.38	(.24,.52)
	<i>incorrect</i>	11	.64	(.34,.94)	.36	(.06,.66)	.18	(-.06,.42)
	<i>correct</i>	39	.44	(.28,.60)	.51	(.35,.67)	.44	(.28,.60)

Table 39: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 4.2.c1

ELEMENT 1								
Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.3_5a	<i>all items</i>	51	.37	(.24,.50)	.43	(.29,.57)	.31	(.18,.44)
	<i>incorrect</i>	14	.21	(-.01,.43)	.21	(-.01,.43)	.36	(.10,.62)
	<i>correct</i>	37	.43	(.27,.59)	.51	(.35,.67)	.30	(.15,.45)
4.3_10a	<i>all items</i>	51	.33	(.20,.46)	.27	(.15,.39)	.31	(.18,.44)
	<i>incorrect</i>	12	.25	(-.01,.51)	.17	(-.05,.39)	.00	(.00,.00)
	<i>correct</i>	39	.36	(.21,.51)	.31	(.16,.46)	.41	(.25,.57)
4.3_17a	<i>all items</i>	52	.33	(.20,.46)	.23	(.11,.35)	.21	(.10,.32)
	<i>incorrect</i>	19	.37	(.15,.59)	.11	(-.03,.25)	.21	(.02,.40)
	<i>correct</i>	33	.30	(.14,.46)	.30	(.14,.46)	.21	(.07,.35)
4.3_18a	<i>all items</i>	51	.39	(.25,.53)	.27	(.15,.39)	.25	(.13,.37)
	<i>incorrect</i>	14	.29	(.04,.54)	.36	(.10,.62)	.21	(-.01,.43)
	<i>correct</i>	37	.43	(.27,.59)	.24	(.10,.38)	.27	(.12,.42)
4.3_23a	<i>all items</i>	52	.08	(.01,.15)	.19	(.08,.30)	.00	(.00,.00)
	<i>incorrect</i>	12	.08	(-.08,.24)	.08	(-.08,.24)	.00	(.00,.00)
	<i>correct</i>	40	.08	(-.01,.16)	.22	(.10,.36)	.00	(.00,.00)

ELEMENT 2								
Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.3_5a	<i>all items</i>	51	.49	(.35,.63)	.53	(.39,.67)	.24	(.12,.36)
	<i>incorrect</i>	14	.36	(.10,.62)	.43	(.16,.70)	.14	(-.05,.33)
	<i>correct</i>	37	.54	(.38,.70)	.57	(.41,.73)	.27	(.12,.42)
4.3_10a	<i>all items</i>	51	.35	(.22,.48)	.55	(.41,.69)	.35	(.22,.48)
	<i>incorrect</i>	12	.25	(-.01,.51)	.33	(.05,.61)	.25	(-.01,.51)
	<i>correct</i>	39	.38	(.23,.53)	.62	(.47,.77)	.38	(.23,.53)
4.3_17a	<i>all items</i>	52	.37	(.24,.50)	.54	(.40,.68)	.52	(.38,.66)
	<i>incorrect</i>	19	.26	(.06,.46)	.47	(.24,.70)	.26	(.06,.46)
	<i>correct</i>	33	.42	(.25,.59)	.58	(.41,.75)	.67	(.51,.83)
4.3_18a	<i>all items</i>	51	.14	(.04,.24)	.73	(.61,.85)	.06	(-.01,.13)
	<i>incorrect</i>	14	.14	(-.05,.33)	.43	(.16,.70)	.07	(-.07,.21)
	<i>correct</i>	37	.14	(.03,.25)	.84	(.72,.96)	.05	(-.02,.12)
4.3_23a	<i>all items</i>	52	.27	(.15,.39)	.67	(.54,.80)	.63	(.50,.76)
	<i>incorrect</i>	12	.25	(-.01,.51)	.50	(.20,.80)	.50	(.20,.80)
	<i>correct</i>	40	.28	(.14,.42)	.72	(.58,.86)	.68	(.52,.82)

Table 40: Mean (95%-confidence interval) of the relative frequency of the analyzed bi-grams for each possible item of Level 4.2.c2

ELEMENT 1								
Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.3_6a	<i>all items</i>	52	.40	(.27,.53)	.46	(.32,.60)	.37	(.24,.50)
	<i>incorrect</i>	31	.32	(.15,.49)	.45	(.27,.63)	.35	(.18,.52)
	<i>correct</i>	21	.52	(.30,.74)	.48	(.26,.70)	.38	(.17,.59)
4.3_8a	<i>all items</i>	51	.31	(.18,.44)	.45	(.31,.59)	.20	(.09,.31)
	<i>incorrect</i>	30	.37	(.19,.55)	.50	(.32,.68)	.23	(.08,.38)
	<i>correct</i>	21	.24	(.05,.43)	.38	(.17,.59)	.14	(-.01,.29)
4.3_14a	<i>all items</i>	51	.33	(.20,.46)	.49	(.35,.63)	.22	(.11,.33)
	<i>incorrect</i>	32	.34	(.17,.51)	.56	(.39,.73)	.25	(.10,.40)
	<i>correct</i>	19	.32	(.11,.53)	.37	(.15,.59)	.16	(-.01,.33)
4.3_16a	<i>all items</i>	52	.21	(.10,.32)	.44	(.30,.58)	.15	(.05,.25)
	<i>incorrect</i>	32	.19	(.05,.33)	.44	(.27,.61)	.12	(.01,.24)
	<i>correct</i>	20	.25	(.06,.44)	.45	(.23,.67)	.20	(.02,.38)
4.3_22a	<i>all items</i>	51	.33	(.20,.46)	.47	(.33,.61)	.20	(.09,.31)
	<i>incorrect</i>	29	.28	(.11,.45)	.38	(.20,.56)	.14	(.01,.27)
	<i>correct</i>	22	.41	(.20,.62)	.59	(.38,.80)	.27	(.08,.46)

ELEMENT 2

Item	Item sample	N	Comparison I		Comparison II		Rotation	
4.3_6a	<i>all items</i>	52	.38	(.25,.51)	.60	(.47,.73)	.44	(.30,.58)
	<i>incorrect</i>	31	.26	(.10,.42)	.65	(.48,.82)	.35	(.18,.52)
	<i>correct</i>	21	.57	(.35,.79)	.52	(.30,.74)	.57	(.35,.79)
4.3_8a	<i>all items</i>	51	.37	(.24,.50)	.25	(.13,.37)	.10	(.02,.18)
	<i>incorrect</i>	30	.30	(.13,.47)	.20	(.05,.35)	.13	(.01,.25)
	<i>correct</i>	21	.48	(.26,.70)	.33	(.12,.54)	.05	(-.04,.14)
4.3_14a	<i>all items</i>	51	.41	(.27,.55)	.49	(.35,.63)	.22	(.11,.33)
	<i>incorrect</i>	32	.34	(.17,.51)	.38	(.20,.55)	.16	(.03,.29)
	<i>correct</i>	19	.53	(.30,.76)	.68	(.47,.89)	.32	(.11,.53)
4.3_16a	<i>all items</i>	52	.52	(.38,.66)	.33	(.20,.46)	.12	(.03,.21)
	<i>incorrect</i>	32	.41	(.24,.58)	.34	(.17,.51)	.16	(.03,.29)
	<i>correct</i>	20	.70	(.49,.91)	.30	(.09,.51)	.05	(-.05,.15)
4.3_22a	<i>all items</i>	51	.65	(.52,.78)	.45	(.31,.59)	.39	(.25,.53)
	<i>incorrect</i>	29	.62	(.44,.80)	.38	(.20,.56)	.34	(.16,.52)
	<i>correct</i>	22	.68	(.48,.88)	.55	(.34,.76)	.45	(.24,.66)