

The interaction of task complexity and gender on mental rotation skills in primary school students

Dinah Reuter^{1*} , Frank Reinhold¹ 

¹Institute for Mathematics Education, University of Education Freiburg, Freiburg, GERMANY

*Corresponding Author: dinah.reuter@ph-freiburg.de

Citation: Reuter, D., & Reinhold, F. (2025). The interaction of task complexity and gender on mental rotation skills in primary school students. *International Electronic Journal of Mathematics Education*, 20(3), em0834. <https://doi.org/10.29333/iejme/16234>

ARTICLE INFO

Received: 19 Dec. 2024

Accepted: 14 Mar. 2025

ABSTRACT

The perception and manipulation of spatial information are essential for mathematical learning, and research highlights gender differences in spatial abilities. The present study contributes to the question of whether these differences are evident at earlier ages and how they interact with task complexity in mental rotation. We developed the 'MEntal Reflection and ROTation' (MERRO) test for group testing in elementary schools, distinguishing between difficulty- and complexity-generating factors, such as context, dimensionality, rotation angle, and number of rotation axes. A pre-study with 148 secondary school students validated the MERRO as suitable for assessing mental rotation skills with another available instrument appropriate for that age group. The main study, involving 387 primary school students (grade 1-grade 4), identified difficulty-generating and complexity-generating factors and investigated potential early gender differences in spatial abilities. Results revealed that stimulus characteristics significantly impact task complexity, and that these characteristics partly interact with gender. These findings offer new insights into the nuanced development of spatial abilities in early mathematics education.

Keywords: mental rotation, gender, primary school, mathematics education, spatial ability

INTRODUCTION

"The perception of the three-dimensional space around us and the ability to orient ourselves in space or to operate mentally with spatial events are human qualifications of practical importance in life" (Franke & Reinhold, 2016, p. 39; translated by author). This high relevance is reflected in a large number of studies—although found in the research tradition of mathematics education: spatial ability is considered one of the most studied cognitive abilities (Maier, 1999, p. 31)—including investigations regarding spatial ability in general and in connection with other cognitive abilities (Battista, 1990; Gilligan-Lee et al., 2022; Guay & McDaniel, 1977; Hawes et al., 2022; Linn & Petersen, 1985; Reinhold et al., 2020; Tam et al., 2019; Thurstone, 1950).

Spatial ability is particularly relevant to mathematics learning and achievement (Gilligan-Lee et al., 2022; see also the SI on "The relation between mathematics achievement and spatial reasoning" by Resnick et al., 2020), high correlations between spatial ability and mathematical skills are commonly found (Hawes et al., 2022; Mix et al., 2016; Smith, 1964; Xie et al., 2020), and studies that show that spatial abilities can be trained also show transfer effects on mathematics achievement in some cases (Cheng & Mix, 2014; Hawes et al., 2017).

Moreover, gender effects in spatial abilities are among the most prominent findings in educational research, while there are no gender differences for general intelligence (Chen et al., 2015). Such gender disparities regarding spatial ability are most prominently found in favour for males—from the age of 10 years onwards (Hoyek et al., 2012; Linn & Petersen, 1985; Reinhold et al., 2020). Yet, one relevant and rather unanswered question is

- (1) if those differences can be found at earlier ages and
- (2) if those differences interact with the complexity of the (mental rotation) task.

For that, we developed an instrument aiming at

- (1) a valid instrument for group tests in elementary schools from grade 1 to grade 4 and
- (2) differentiating between a broad variety of difficulty generating factors in mental rotation tasks (contextualized or abstract, plain or three-dimensional-perspective, rotation angle, number of rotation axis).

Spatial Ability, Gender, and Mathematics Performance

Spatial ability is considered a fundamental cognitive skill that represents a sub-component of human intelligence (Gardner, 1983; Linn & Petersen, 1985; Thurstone, 1950). Despite various studies using a wide variety of terms to frame spatial ability, as well as different underlying conceptualizations, there seems to be an agreed upon general definition (Linn & Petersen, 1985; Maier, 1999; Thurstone, 1950); based on a meta-analysis, Linn and Petersen (1985) divide spatial ability into three sub-areas: spatial perception, mental rotation, and spatial visualization. These sub-areas include requirements that are related to one's own person in relation to a spatial position (i.e., spatial perception) as well as requirements that are independent of this (mental rotation and spatial visualization; Maier, 1999).

Above that, spatial ability has a high practical relevance in everyday interaction with the environment (Barke, 1980; Bodner & Guay, 1997; Freudenthal, 1971). Most prominently, spatial ability is considered to be of importance with regard to dealing with mathematical content (Battista, 1990; Besuden, 1999; Lorenz, 1992; Maier, 1999). Decades of empirical research show a high correlation between spatial ability and mathematical skills (Bodner & Guay, 1997; Gilligan-Lee et al., 2022; Grüßing, 2012; Hawes et al., 2022; Mix et al., 2016; Plath, 2014; Smith, 1964; Xie et al., 2020)–both with regard to general mathematical skills (Bodner & Guay, 1997; Linn & Petersen, 1985) as well as with regard to specific content areas such as geometry (Battista, 1990) or arithmetic (Gunderson et al., 2012; Kajda & Kajda, 2010; Landerl et al., 2022; Tam et al., 2019). In addition, studies have shown that it is possible to foster one's spatial ability and that this can also have a positive impact on mathematical performance (Adaboh et al., 2017; Cheng & Mix, 2014; Hawes et al., 2017; Nes & Doorman, 2011; Uttal et al., 2013; Verdine et al., 2017). However, there are indications that the way the fostering is provided has a gender-specific effect: study results suggest that boys benefit more from a highly guided situation than girls (Weber et al., 2024).

While there are controversial discussions with regard to gender differences for various cognitive abilities due to inconclusive research results, the empirical situation in relation to spatial ability and gender is undisputed (Reilly et al., 2017). Various studies confirm gender differences in the area of spatial ability in favor of male participants (Bartlett & Camba, 2023; Levine et al., 1999; Linn & Petersen, 1985; Reinhold et al., 2020). Several explanations for the observed differences are discussed, which include social, biological, and cultural influences (Reilly et al., 2017). Furthermore, gender-stereotypical beliefs can already be detected in preschool children with regard to activities that address spatial abilities (Ebert et al., 2024).

Mental Rotation as a Sub-Area of Spatial Ability

Mental rotation includes the ability to rotate two- and three-dimensional objects in one's mind (Shepard & Metzler, 1971). Standardized tests such as the mental rotation test (MRT) (Peters et al., 1995; Vandenberg & Kuse, 1978) for (young) adults and modified, simpler versions for younger age groups (Jansen et al., 2013; Levine et al., 1999) are commonly used to assess this ability. In these tests, participants are asked to complete as many items as possible without errors in a given time. Typically, participants must decide whether the figures shown are rotated or mirrored variants of an original figure. In the instruments commonly used with adults, these figures are usually three-dimensional cube figures, such as those contained in the MRT, or representations based on them (e.g., Rahe et al., 2021); some studies also integrate illustrations of everyday objects in a perspective view (e.g., Rahe et al., 2021; Ruthsatz et al., 2019). Moreover, while in the standardized tests for adults there is typically more than one comparison item to choose from, and only the rotated variants for an initial figure have to be identified, standardized tests for (young) children may also contain only one comparison item for the initial figure, e.g., when the decision has to be made whether it is mirrored or rotated (e.g., Jansen et al., 2013). In general, studies have shown that tasks become more error-prone with increasing angular disparity (Shepard & Metzler, 1971), and it is commonly acknowledged that mental three-dimensional-rotation tests are reliable instruments for assessing spatial ability (Hawes et al., 2015).

For the rotation of two-dimensional objects, studies show that this ability develops from around the age of five and that even younger children are able to successfully complete mental rotation tasks (Frick et al., 2013; Jansen et al., 2013; Levine et al., 1999; Marmor, 1975). The quality of the instructions and familiarity with the objects, especially realistic rotation objects, appear to have a strong influence on the success rate (Jansen et al., 2013). Regarding the ability to mentally rotate three-dimensional objects in younger children, however, the study situation is not quite as clear. Some research findings suggest that the cognitive demands of tasks involving the mental rotation of three-dimensional objects are too high and that younger children are therefore unable to process them successfully (Hoyek et al., 2012; Jansen et al., 2013). Hawes et al. (2015) identified difficulty generating factors that can hinder successful processing–regardless of the ability to mentally rotate. These include the challenge that, in a conventional paper-pencil test, the two-dimensional representation of a three-dimensional object must first be mentally translated back into three dimensions in order to be able to perform the mental rotation. As a result, Hawes et al. (2015) developed a test instrument with tasks for the mental rotation of three-dimensional objects with lower cognitive requirements and were thus able to successfully demonstrate the development of the mental rotation of three-dimensional objects in the first years of primary school and, thus, the ability to mentally rotate three-dimensional objects in younger children.

Studies on mental rotation show gender differences in favor of male participants (Levine et al., 2016; Linn & Petersen, 1985; Peters et al., 1995, 2006; Voyer et al., 1995). However, prominent research from German speaking research traditions shows stimulus-dependent gender differences. For example, the classic cube figures from the MRT appear to be easier for males than for females, while no significant gender differences were found for the comparable ball items (Rahe & Quaiser-Pohl, 2019). Other studies also suggest that no gender differences emerge for items with no or less gender-stereotypical attributions (Lennon-Maslin & Quaiser-Pohl, 2024).

In addition to gender disparities in mere performance, studies that focus on the choice of strategy have also shown that there are differences in the mental comparison process between men and women: while men tend to use a general, holistic strategy, women tend to take a more analytical approach–in the sense of a 'piece-by-piece rotation' (Cochran & Wheatley, 1989; Corballis,

1997). Moreover, studies with younger participants confirm gender-specific differences both with regard to the success rate from the age of ten (Hoyek et al., 2012; Linn & Petersen, 1985), and with regard to the differences in the choice of strategy in children aged eleven and over (Geiser et al., 2008). For children of primary school age, studies have shown that gender-specific differences in the area of mental rotation cannot yet be (reliably) mapped (Bott et al., 2023; Grüßing, 2012; McGee, 1979). Some studies confirm stimulus-dependent gender differences for this age group, e.g., gender differences with stimuli that are very close to everyday life, which can also be categorized according to gender stereotypes. Here, studies show that it is easier for children to mentally rotate stimuli that can be considered congruent with typical gender stereotypes (Ruthsatz et al., 2015, 2019).

The Present Studies

In summary, studies regarding mental rotation skills show gender differences in terms of success rate as well as stimulus-dependent differences in terms of proximity to everyday life and gender stereotypes. At the same time, we argue that there is a lack of item consistency: different studies use different items and do usually not provide comprehensive item analyses. For example, there are discussions about the construction of the items with regard to closeness to and distance from everyday life as well as gender stereotypes—but items also differ with regard to non-perspective and perspective two-dimensional representations of three-dimensional objects, with regard to two- and three-dimensional rotation, and with regard to the basic complexity of the object. Therefore, we argue that it is plausible and relevant to control for item characteristics that may affect the empirical ease of an item, i.e., *difficulty* generating factors of the *item* (e.g., rotation angle) and *complexity* generating factors of the *stimulus* (e.g., context and/or perspective) systematically during item development. For that, we developed the 'MEntal Reflection and ROTation' test (MERRO)—an instrument aiming at differentiating between a broad variety of difficulty generating factors in mental rotation tasks, which is suitable for use in primary school.

In a pre-study, we examined the validity of the new test instrument by comparing it to the MRT (Peters et al., 1995) and explored indications of gender differences compared to the established measure for a cohort of teenagers. As there are up to our knowledge no comparable group tests for primary school children available, we conducted pre-study with teenagers using an established measure for mental rotation skills as a benchmark. This approach allowed us to assess whether the newly derived measure captures the same underlying construct as the MRT before applying it to a younger cohort (where up to our knowledge no validated group test exists). This led to the following research question (RQ) which will be answered along with two sub-questions:

RQ1. *Is the MERRO test a valid instrument to assess mental rotation skills—utilizing the MRT as a commonly-agreed upon operationalization of mental rotation skills?*

More specifically:

RQ1a. *To what extent does the MERRO demonstrate sensitivity to gender differences, in comparison to the (potential) gender differences in the same group of participants when assessed with the MRT?*

RQ1b. *Is there a substantial correlation between the MERRO and the MRT, and does gender affect this correlation?*

In the main study, we used the newly-designed MERRO test instrument in various primary school classes to check which levels of empirical difficulty can be found in the test items, and to what extent gender differences can already be found in this age group—leading to the following theory-driven explorative RQs:

RQ2. *Does the number of completed MERRO items in a given time frame vary by age and gender of primary school students?*

RQ3. *How does the solution probability in MERRO items vary with item difficulty, stimulus complexity, and student characteristics?*

More specifically:

RQ3a. *How do grade level, difficulty factors (e.g., rotation angle), and complexity factors (e.g., context and perspective) influence the solution probability of test items among primary school students?*

RQ3b. *What are the effects of gender on solution probabilities, and how do interaction of gender with other factors (e.g., item difficulty and grade level) impact performance in the MERRO in primary school students?*

METHOD AND STUDY DESIGN

This study employs a quantitative, cross-sectional research design to investigate the relationship between task complexity and gender differences in mental rotation skills among primary school students. It consists of two distinct phases:

- (1) a pre-study, which validates our newly developed MERRO with the well-established MRT (Peters et al., 1995) and
- (2) the main study, which applies the MERRO to a large sample of primary school students (grade 1-grade 4).

The pre-study was conducted with an older sample (secondary school students) because the MRT is designed and validated for (young) adults and older students. Since the MRT serves as a widely accepted measure of mental rotation ability, it was crucial to establish whether the MERRO captures the same underlying construct. This validation step ensures that the MERRO aligns with the standardized theoretical framework of mental rotation—while being adapted for younger learners.

The main study, in contrast, investigates how task complexity interacts with gender differences in mental rotation performance within the target population—primary school students. Statistical analyses assess the role of stimulus characteristics in shaping mental rotation ability at an early age.

Table 1. Participants from main study per grade level, the respective gender ratio as well as the mean (*M*) age of the subgroups

	Grade 1		Grade 2		Grade 3		Grade 4	
	<i>N</i>	<i>M (age)</i>	<i>N</i>	<i>M (age)</i>	<i>N</i>	<i>M (age)</i>	<i>N</i>	<i>M (age)</i>
Total	159	7.013	64	7.984	94	8.968	70	9.871
Male	79	7.076	38	7.974	45	9.000	38	9.921
Female	80	6.950	26	8.000	49	8.939	32	9.813

Note. Age was assessed by asking the students how old they were at the time of the assessment and answers were given in years only

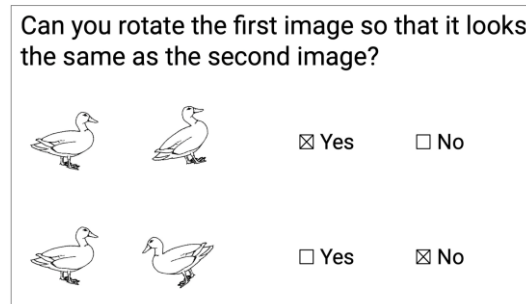


Figure 1. Typical stimulus in the MERRO test—depicted with the training item which is only rotated for 25 degrees (Source: Authors' own elaboration)

Sample

Both samples were collected after the local school authority evaluated and approved the study (approval number 7-6499.2). The sample in the pre-study consisted of $N = 148$ secondary school students ($n = 82$ boys and $n = 56$ girls, and $n = 10$ students did not give an answer) with an M age of 15.6 years (*standard deviation* [SD] = 0.68). All children gave verbal informed consent to take part in the study, after their parents gave written informed consent. The sample in the main study consisted of $N = 387$ primary school students ($n = 200$ boys and $n = 187$ girls) from grade 1 to grade 4—nested in $k = 24$ classrooms. The distribution of all participants across the four grade levels, the gender ratio per grade level and the M age of the participants is depicted in **Table 1**. All children gave verbal informed consent to take part in the study, after their parents gave written informed consent.

Instrument

In line with our RQs and hypotheses, we developed a complex itemset for assessing spatial ability in primary school students via mental rotation. Our aim was to include general difficulty generating factors (rotation angle) as well as a variety of complexity increasing factors (context of the pictorial representations; plain vs. presentations in perspective; number of bends, tilts in cube stimuli) to get a detailed picture of potential gender differences in this age group.

The itemset was finalized after three piloting cycles; we first conducted qualitative piloting with $N = 4$ students who were closely monitored during test-taking and interviewed informally after the test; a first quantitative pilot study with $N = 33$ students was used to select the items that had to be altered because of problems with the stimuli; a second quantitative pilot study with $N = 13$ students was utilized to get a first impression of retest reliability.

The MERRO test uses 24 different stimuli created from 10 different figures. It assesses students' ability to decide whether the stimuli were (a) only rotated in the paper plain, or (b) mirrored on the y -axis and then rotated in the paper plain. For that, we use the age-appropriate question "Can you rotate the first image so that it looks the same as the second image?" for all items, which students have to answer on a "yes" and "no" basis (**Figure 1**).

The 10 figures (**Figure 2**, separate lines) differ in four contexts of varying theoretical complexity: In context 0, authentic pictures are used (**Figure 2**, stimuli 021, 209, 307, 308, 309, 407, 408, and 409); in context 1, cube images with one or two bends are used (**Figure 2**, stimuli 301, 302, 303, 401, 402, and 403); in context 2, cube images with three planar bends are used (**Figure 2**, stimuli 303, 304, 305, 403, 404, and 405); in context 3, cube images with three twisted bends are used (**Figure 2**, stimuli 405, 406, 505, and 506). The stimuli are either presented plain (**Figure 2**, first column), or in perspective (**Figure 2**, second and third column). Among the ones presented in perspective, the stimuli differ in whether they are tilted in z -direction (**Figure 2**, third column), or not (**Figure 2**, second column). Figure 021 and figure 209 were copied from Snodgrass and Vanderwart (1980), figure 307 and figure 308 were copied from Google Sketchup, all other figures were developed by the authors in order to fit into the displayed contexts; items 405 and 406 are inspired by the MRT (Peters et al., 1995).

To ensure a comprehensive assessment of spatial ability at different levels of elaboration, complex items (e.g., 405, 406, 505, and 506) were intentionally included. We argue that these items will allow them to capture a wide range of spatial reasoning skills and provide a detailed differentiation of students' abilities rather than merely identifying success rates on simpler rotations.

In addition to the aforementioned complexity increasing factors (context, in perspective, and tilted), difficulty is further varied by increasing the rotating angle (45, 90, and 135 degrees).

Instruments used in pre-study

In the pre-study, we used four comparable short versions of the MERRO test. Each of them included all 24 stimuli depicted in **Figure 2**.

























Context	Plain	In perspective		
		Not tilted	Tilted	
Context 0: Authentic pictures				
	Stimulus 021			
				
	Stimulus 209			
				
	Stimulus 307	Stimulus 407	Stimulus 507	
				
Stimulus 308	Stimulus 408	Stimulus 508		
Context 1: Cube images with one or two bends				
	Stimulus 301	Stimulus 401	Stimulus 501	
				
	Stimulus 302	Stimulus 402	Stimulus 502	
	Context 2: Cube images with three bends			
		Stimulus 303	Stimulus 403	Stimulus 503
				
Stimulus 304		Stimulus 404	Stimulus 504	
Context 3: Cube images with three twisted bends				
			Stimulus 405	Stimulus 505
				
		Stimulus 406	Stimulus 506	

Figure 2. Overview of the 24 stimuli made of 10 figures used in the MERRO test (Source: Authors' own elaboration)

The four versions were blocked with variation in the rotation angle—each block containing a balanced number of rotations of 45 degrees, 90 degrees, and 135 degrees (8 each). Of the 24 items presented, either 11, 12, or 13 were mirrored and rotated in the four booklets. All four booklets yielded good reliability, with an *M* Cronbach's $\alpha = 0.84$.

The other measure used to assess spatial ability in pre-study was the MRT by Peters et al. (1995), which is based on the original paper-pencil test by Vandenberg and Kuse (1978). The test consists of 24 items. In each item, three-dimensional and complex (bend and tilted) geometrical structures built out of cubes is given. Two correctly rotated structures corresponding to the initial structure must be picked out of a selection of four. We utilized two different ways of coding the MRT. In the “hard” coding, full credit is given only when both correct structures are marked, as proposed by Peters et al. (1995, 2006). The “hard” coding yielded a very good reliability, Cronbach's $\alpha = 0.90$. As the MERRO test—which we aimed to compare the results of the MRT to—is considerably easier than the MRT, we also used a “easy” coding of the MRT, where partial credit is given for one out of two correctly marked structures. The “easy” coding also yielded a very good reliability, Cronbach's $\alpha = 0.92$.

In both assessments in the pre-study, the MERRO and the MRT, a total of 24 points could possibly be achieved.

Instruments used in the main study

In the main study, the MERRO test consisted of 48 items (2 for each of the 24 stimuli), balanced with three degrees of rotation, with 24 items being mirrored and rotated, and 24 items only being rotated.

Procedure

Procedure of the pre-study

The study was first approved by the local school authority, the students as well as their parents gave informed consent to take part in the survey, and the assessment was conducted afterwards during the regular mathematics lessons in the children's classrooms.

The short version of the MERRO was administered as a group test in paper-based format with a 6-page DIN-A4 test booklet. The first page showed the introductory item (**Figure 1**), and students were guided in marking the correct responses. The second page showed a 'stop' sign, marking the start of the timed assessment. On the following 4 pages, 6 different items were displayed in randomized but identical order. The students had 2 minutes to complete as many items as possible—the timing was introduced by the phrase "Solve as many tasks as possible. How many tasks can you complete in 2 minutes?" One version of the four booklets was randomly assigned to each student taking part.

The MRT was also administered in paper-based format, with a standardized introductory part introducing the assessment (Peters et al., 2006). After students confirmed that they understood how to proceed in the MRT, they had 3 + 3 minutes for answering the 24 items (with a short break between the first and the second 12 items, as suggested).

The test booklets were human coded by noting "yes", "no", and "NA" at the respective item of the MERRO test, and by noting two-digit codes corresponding to the students' selections (e.g., the code "24" would represent that a student chose the second and the fourth figure, "39" a selection of only the third figure, and "9999" no selection in the respective task). To avoid mistakes, correctness of the solution was derived by applying a solution vector to the "yes" and "no" coding as part of the analysis.

Procedure of the main study

After the main study was approved by the local school authority and the students as well as their parents gave informed consent to take part in the survey, the assessment was conducted during the regular mathematics lessons in the children's classrooms.

The MERRO test was administered exactly like in the pre-study (group test in paper-based format, 10-page DIN-A4 test booklet). Following the introductory page (**Figure 1**) and the second page that marked the timed assessment, on each of the 8 pages, 6 different items were displayed in randomized but identical order. The students had 4 minutes to complete as many items as possible—the timing was introduced by the phrase "Solve as many tasks as possible. How many tasks can you complete in 4 minutes?"

After the assessment, the test booklets were human coded by noting "yes", "no", and "NA" at the respective item. Again, correctness of the solution was derived by applying a solution vector to the coding as part of the analysis.

Analysis

All data preparation and analysis were conducted in *R* (R Core Team, 2008).

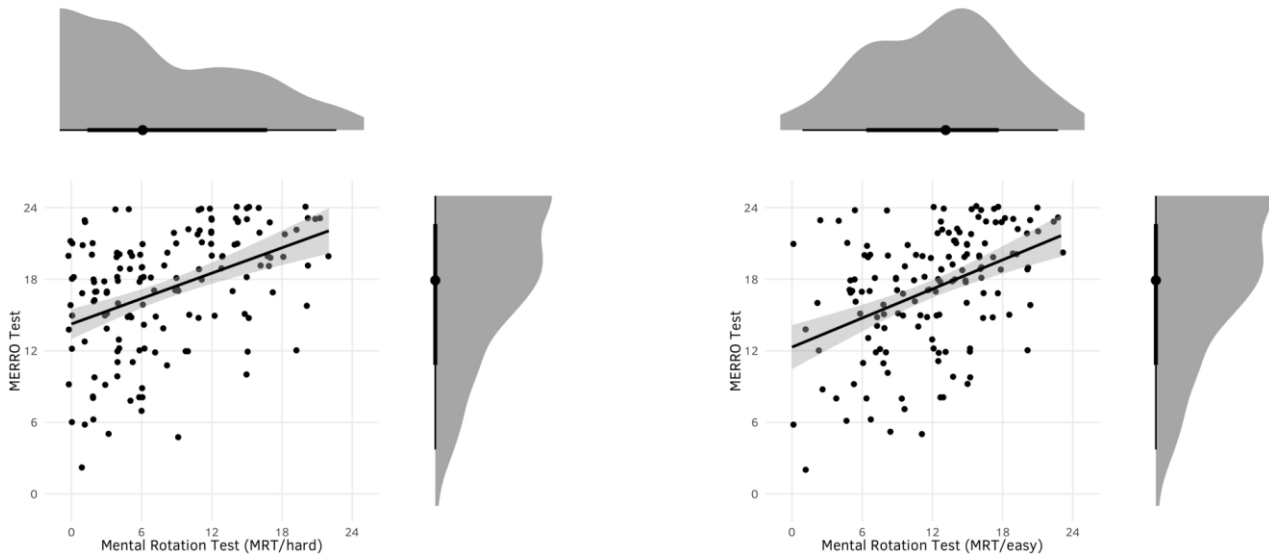
In the pre-study, we estimated gender differences in both instruments using Welsh *t*-tests and Cohen's *d* measures of the effect sizes (**RQ1a**). Furthermore, we calculated bivariate Pearson correlations for the total sample—and the gender-specific subsamples (**RQ1b**). In the pre-study, we were interested in person characteristics (not item characteristics, as compared to the main study), i.e., all "NAs" from the timed tests were counted as "wrong"—as sum scores were compared.

For **RQ2** (main study), we utilized a Poisson generalized linear mixed models (GLMM) to estimate the effects of grade level and gender on the number of completed tasks. Mixed models were estimated using the `{lme4}`-package (Bates et al., 2015). All models allowed for a classroom random intercept. Model 0 did not contain any fixed effects. Model 1 contained only grade level (ordered factor, only linear term was estimated) as a fixed effect. Model 2 also contained gender (0 = male and 1 = female) and the interaction effect of grade level and gender as fixed effects. Effect sizes (and their confidence intervals) are given as incidence rate ratios (*IRR*), describing the relative difference between the number of completed tasks when a predictor increases by 1 compared to the baseline (with an *IRR* between 0 and 1 describing a decline in the number of completed tasks, and an *IRR* > 1 describing an increase in the number of completed tasks).

For **RQ3** (main study), we utilized binomial GLMMs to estimate the effects of grade level, difficulty generating factors (degree rotated) and all complexity increasing factors (context, in perspective, tilted), as well as their interaction with gender on the solution probability. All models allowed for a student and a classroom random intercept (students nested in classrooms), as well as a stimulus and figure random intercept (stimuli nested in figures). As we were mainly interested in item characteristics, "NAs" were ignored for this analysis (and not coded as "incorrect"). Again, model 0 did not contain any fixed effects. Model 1 contained the following fixed effects: Contexts (see **Figure 2**) were differentiated by three contrasts: *cubes* (-0.5 = context 1 and 0.5 = contexts 2, 3, & 4); *bends* (0 = context 1, -0.5 = context 2, and 0.5 = context 3 & context 4); *twisted* (0 = context 1 & context 2, -0.5 = context 3, and 0.5 = context 4). The remaining complexity increasing factors were also differentiated by two contrasts: *in perspective* (-0.5 = column 1 and 0.5 = column 2 & column 3 in **Figure 2**); *tilted* (0 = column 1, -0.5 = column 2, and 0.5 = column 3 in **Figure 2**). The *degrees rotated* and *grade level* were estimated as ordered factors, where only the linear term was used in the models. Model 2 also contained gender (0 = male and 1 = female) and the interaction effects of gender and all other predictors as fixed effects. For the ease of interpretation, all predictors are displayed in meaningful values in the following figures in the results section: Effect sizes (and their confidence intervals) are given as odds ratios (*OR*), which are an indicator of the multiplicative change in the solution probability per change of the binary predictor variables from 0 to 1—or per increase of 1 in the ordinal predictor variables (with an *OR* between 0 and 1 describing a decline in the solution probability, and an *OR* > 1 describing an increase in the solution probability).

Table 2. Gender differences in the MERRO test and the MRT (validation pre-study)

	Total		Male		Female		Gender differences			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
MERRO test	17.128	5.019	17.537	4.879	16.714	5.169	0.939	113.71	0.350	0.165
MRT (hard)	8.108	5.917	9.098	6.510	6.679	4.940	2.478	134.43	0.014	0.408
MRT (easy)	11.858	5.219	12.811	5.396	10.661	4.929	2.421	125.02	0.017	0.413

**Figure 3.** Correlation between the MERRO test and the MRT, scatterplots, and distribution in the instruments (validation pre-study) (Source: Authors' own elaboration)

In line with our research interest, we focus on the interpretation of effect sizes (not statistical significance), as they provide a more direct measure of the magnitude of the effects of interest, as well as their interpretation. As an additional measure of the size of the effects of the predictors in GLMMs (besides the *IRRs* and the *ORs*), the *proportion change in variance* (PCV) (Nakagawa & Schielzeth, 2013) is given, which describes a change in a random intercept between two models when additional fixed effects are considered.

RESULTS

Pre-Study: Assessment of Mental Rotation Skills

Our first RQ aimed at validating the developed MERRO test with an established instrument for the assessment of spatial ability with mental rotation items. We chose the MRT (Peters et al., 2006)–which is developed for an older age group than primary school students–which is why a sample of teenagers had to be assessed. For this validation, we firstly estimated gender differences in the MRT and the MERRO test in the same sample (**RQ1a**), with the hypothesis to find stronger gender differences in the MRT in favor for male students, because the MERRO test also contains items where we would not expect gender differences at all. **Table 2** shows that the results are in line with this assumption. There were significant gender effects in favor for boys in both interpretations of the MRT (hard coding: $d = 0.408$ and easy coding: $d = 0.413$, **Table 2**), while in the same sample the MERRO test showed a non-significant and considerably smaller gender effect in favor for boys, $d = 0.165$.

Lastly, we investigated the correlation between the MERRO test and the MRT (in a sample where both tests could be administered)–and whether these correlations vary with regard to students' gender (**RQ1b**). There was an overall positive correlation between the MERRO test and the MRT (hard coding), $r(146) = 0.419$ and $p < .001$, and the MRT (easy coding), $r(146) = 0.423$ and $p < .001$ (**Figure 3**). This effect was stronger for boys than for girls both in the hard coding of the MRT, $r_{\text{mal}}(54) = 0.481$, $p < .001$, $r_{\text{fem}}(80) = 0.374$, and $p < .001$, and the easy coding of the MRT, $r_{\text{mal}}(54) = 0.458$, $p < .001$, $r_{\text{fem}}(80) = 0.378$, and $p < .001$ (**Figure 4**).

Main Study: The Interaction of Task Complexity and Gender on Mental Rotation Skills in Primary School Students

Number of completed tasks

Besides the mere solution rate, the number of completed tasks–given a restricted time for the assessment–serves as an indicator of spatial ability (**RQ2**). If the ability to rotate images mentally should increase (as expected) by age, students from lower grade levels should solve less items than students from higher grade levels–and a potential gender effect should also result in a difference in the number of completed tasks between boys and girls of a respective grade level.

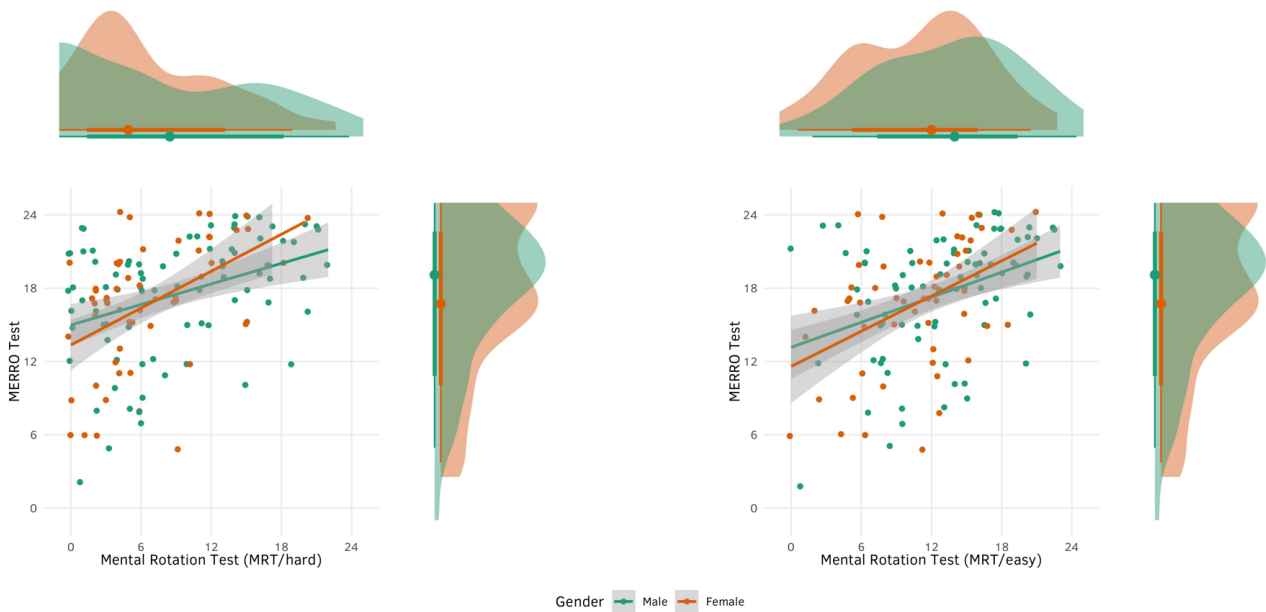


Figure 4. Correlation between the MERRO test and the MRT as a function of gender, scatterplots, and distribution in the instruments (validation pre-study) (Source: Authors' own elaboration)

Table 3. Number of completed tasks as a function of grade level and gender (main study)

Fixed effects	Model 0				Model 1				Model 2			
	IRR	SE	95% CI		IRR	SE	95% CI		IRR	SE	95% CI	
			LL	UL			LL	UL			LL	UL
Intercept	28.597	1.245	26.258	31.144	24.431	1.130	22.313	26.750	25.577	1.257	23.229	28.162
Grade level ^a					1.140	0.032	1.079	1.204	1.150	0.034	1.086	1.219
Gender ^b									0.907	0.027	0.856	0.961
× Grade level									0.981	0.016	0.949	1.013
Random effects	Variance	SD			Variance	SD			Variance	SD		
Classroom	0.043	0.207			0.021	0.145			0.022	0.149		
Model fit	AIC	BIC	R ² (m)	R ² (c)	AIC	BIC	R ² (m)	R ² (c)	AIC	BIC	R ² (m)	R ² (c)
	3,567.3	3,575.2	0.000	0.556	3,553.7	3,565.5	0.293	0.562	3,514.8	3,534.6	0.323	0.588

Note. Number of observations = 387; Number of classrooms = 24; SE: Standard error; CI: Confidence interval; LL: Lower limit; UL: Upper limit; AIC: Akaike information criterion; BIC: Bayesian information criterion; R² (m): Marginal R²; R² (c): Conditional R²; ^a0: Grade 1; 1: Grade 2; 2: Grade 3; & 3: Grade 4; & ^b0: Male & 1: Female

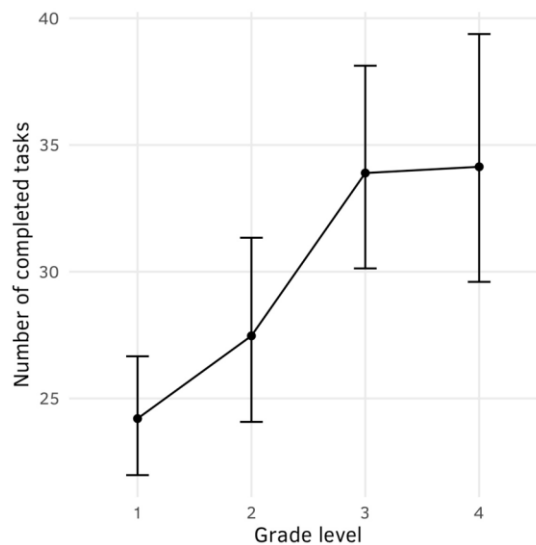


Figure 5. Effect of grade level on the number of completed tasks (model 1 in Table 2, main study) (a total number of 48 completed tasks was possible in this assessment) (Source: Authors' own elaboration)

Table 3 and Figure 5 and Figure 6 show that both can be considered true.

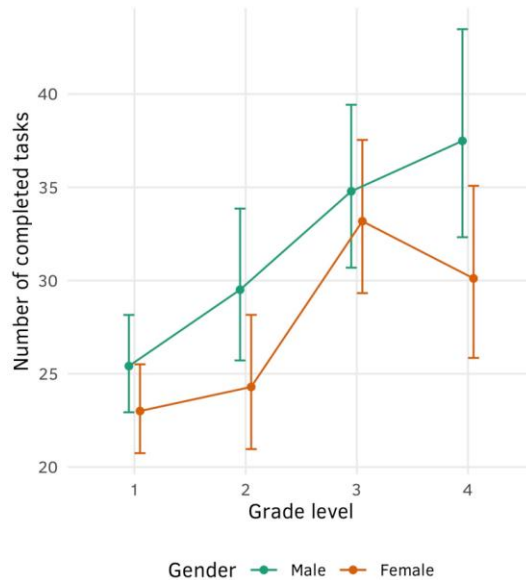


Figure 6. Effects of grade level and gender on the number of completed tasks (model 2 in **Table 2**, main study) (a total number of 48 completed tasks was possible in this assessment) (Source: Authors’ own elaboration)

Table 4. Estimated solution probability as a function of grade level, gender, complexity, and difficulty (main study)

Fixed effects	Model 0				Model 1				Model 2			
	95% CI				95% CI				95% CI			
	OR	SE	LL	UL	OR	SE	LL	UL	OR	SE	LL	UL
Intercept	2.611	0.380	1.963	3.474	3.121	0.420	2.397	4.063	3.290	0.493	2.453	4.412
Cubes ^a					1.016	0.189	0.705	1.463	1.037	0.199	0.711	1.511
Bends ^b					0.724	0.162	0.466	1.124	0.738	0.171	0.469	1.161
Twisted ^c					1.098	0.291	0.653	1.848	1.072	0.294	0.627	1.835
In perspective ^d					0.744	0.081	0.602	0.921	0.756	0.089	0.600	0.953
Tilted ^e					0.726	0.080	0.585	0.900	0.776	0.094	0.612	0.984
Angle ^f					0.687	0.023	0.644	0.734	0.667	0.028	0.614	0.724
Grade level ^g					1.264	0.073	1.128	1.416	1.282	0.086	1.123	1.463
Gender ^h									0.893	0.121	0.684	1.165
× Cubes									0.954	0.094	0.787	1.157
× Bends									0.955	0.108	0.765	1.193
× Twisted									1.049	0.148	0.795	1.384
× In perspective									0.971	0.098	0.796	1.183
× Tilted									0.866	0.096	0.697	1.075
× Angle									1.069	0.061	0.956	1.195
× Grade level									0.970	0.069	0.845	1.114
Random effects	Variance	SD			Variance	SD			Variance	SD		
Students	0.398	0.631			0.405	0.636			0.401	0.634		
In classrooms	0.144	0.380			0.063	0.252			0.065	0.255		
Stimuli	0.098	0.320			0.036	0.189			0.036	0.189		
In figures	0.085	0.292			0.049	0.222			0.050	0.223		
Model fit	AIC	BIC	R² (m)	R² (c)	AIC	BIC	R² (m)	R² (c)	AIC	BIC	R² (m)	R² (c)
	12,958	12,994	0.000	0.181	12,830	12,918	0.054	0.190	12,840	12,987	0.055	0.191

Note. Number of observations = 11,367; Number of students = 387 in 24 classrooms; Number of stimuli = 24 in 10 figures; ^a0: Authentic pictures & 1: Cube images; ^b0: One or two bends & 1: Three bends; ^c0: Plain bends and 1: Twisted bends; ^d0: Plain and 1: In perspective; ^e0: Not tilted and 1: Tilted; ^f0: 45 degrees, 1: 90 degrees, & 2: 135 degrees; ^g0: Grade 1, 1: Grade 2, 2: Grade 3, & 3: Grade 4; & ^h0: Male & 1: Female

The number of completed tasks increases constantly from grade 1 to grade 4 (**Figure 5**), with girls completing less tasks than boys in every grade level (**Figure 6**).

Both the effects of grade level and gender can be considered relevant, with a constant decrease in the AIC and BIC as well as an increase in the marginal R^2 from model 0 to model 2 (**Table 3**). A relevant estimator for the effect of grade level is the PCV from model 0 to model 1: considering the grade level of the students reduces the random variance between classrooms by 52.16%.

Solution probability

Our third RQ distinguished between the effects of age (i.e., grade level), the rotation angle, and the complexity of the stimulus (**RQ3a**), as well as potential differences between boys and girls in these effects (**RQ3b**).

A comprehensive analysis of the main effects yielded results that were largely consistent with our hypotheses (model 1 in **Table 4** and **Figure 7**): The solution probability increased with an increase in students grade level and decreased with an increase

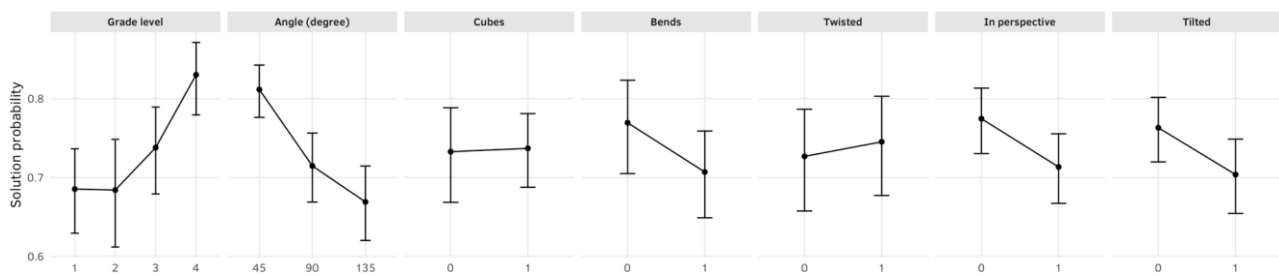


Figure 7. Effects of grade level, difficulty generating, and complexity increasing factors on the solution probability (model 1 in Table 3, main study) (Source: Authors' own elaboration)

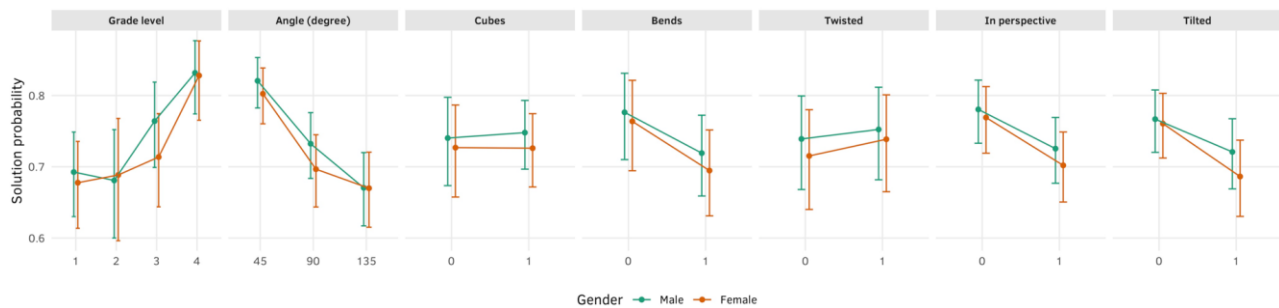


Figure 8. Effects of grade level, difficulty generating, and complexity increasing factors on the solution probability as a function of gender (model 2 in Table 3, main study) (Source: Authors' own elaboration)

in the rotation angle. Regarding the context of the stimuli, there was no notable difference in solution probability for authentic pictures or cube images—when controlled for all other varying item characteristics. The number of bends increased the empirical difficulty of an item, but stimuli with three plain bends were of comparable empirical difficulty to stimuli with three twisted bends. Moreover, items with stimuli displayed in the plain were easier than items with stimuli displayed in perspective; among the stimuli presented in perspective, those tilted in z-direction were more difficult than those not tilted in z-direction. All in all, considering these item-specific and classroom-specific predictors decreased the AIC and BIC—and increased the marginal R^2 —from model 0 to model 1 (Table 4), and led to a PCV of 56.25% in the classroom random intercept, a PCV of 63.27% in the stimulus random intercept, and a PCV of 42.35% in the figure random intercept.

Our hypothesis was that the development of the MERRO test with its complex itemset and item structure would allow for a more fine-grained analysis of gender effects in mental rotation abilities in primary school students. When including gender—and the interactions of all predictors from model 1 with gender—as fixed effects into model 2, a PCV in the students random intercept would indicate an overall gender effect, a PCV in the classroom random intercept would indicate a gender effect interacting with the age of the participants, and PCVs in the stimulus and figure random intercepts would indicate gender differences in how specific complexity increasing factors increase difficulty for boys and girls differently. Our hypothesis was that gender differences in favor of boys would become explicitly visible in items of higher complexity. The results are only partly in line with this hypothesis (model 2 in Table 4 and Figure 8). There was no decrease in the classroom, stimulus, and figure random intercept, and only a marginal PCV of 1.0% in the student random intercept between model 1 and model 2; AIC and BIC did not decrease, and the marginal R^2 increased only slightly to 5.5% (Table 4). A more detailed look at the interaction effects with gender revealed that on a descriptive level, boys outperformed girls in items of every category (Figure 8). Differences in increased empirical difficulty for girls when compared to boys were found in an increased number of bends, when stimuli were presented in perspective, and—most pronounced—in the most complex item types—i.e., items with stimuli displayed in perspective and tilted in z-direction (Figure 8).

DISCUSSION

The MERRO test developed for the present study aims at a valid assessment of mental rotation skills in children at primary school age, synthesizing findings from prior research on younger children's mental rotation skills (e.g., Hawes et al., 2015; Jansen et al., 2013; Levine et al., 1999). It utilizes everyday objects as well as cube figures, both in planar in a perspective view, with increasing levels of complexity. The assessment is implemented by asking students to decide whether figures shown are rotated or mirrored variants of the original figure, while the rotation is done around 45, 90, or 135 degrees only in the paper plain.

The pre-study suggests that the MERRO can be considered a suitable instrument to assess mental rotation skills, compared to a commonly agreed upon measurement—the MRT (Peters et al., 1995). One notable limitation of this study is the use of a teenage sample to validate the newly developed MERRO test. Although this approach was necessary due to the absence of comparable group tests for primary school children, it introduces certain constraints. The cognitive and developmental differences between teenagers and primary school children may affect the generalizability of the findings. Future research should aim to validate the MERRO test directly with primary school children to ensure its suitability and accuracy for this age group. Despite these limitations,

we argue that pre-study provides a foundational validation of the MERRO test, indicating its potential for effective assessment of mental rotation skills in younger children.

The main study did, firstly, replicate findings regarding mental rotation skills in children, such as an overall increased difficulty corresponding to an increasing angular disparity (Shepard & Metzler, 1971), that children from grade 1 to grade 4 are able to successfully complete mental rotation tasks with two dimensional objects (Frick et al., 2013; Jansen et al., 2013; Levine et al., 1999; Marmor, 1975), and that the ability to successfully complete mental rotation tasks increases with age—with the MERRO being sensitive for such development.

Our findings from the main study broaden the perspective on the development of early mental rotation skills. We could show that, indeed, there exist characteristics of the stimulus besides the rotation angle that increase complexity, as suggested also by, e.g., Hawes et al. (2015). Our results suggest that complexity does not increase when switching from everyday objects to cube figures in general, but that also cube figures of low complexity can be generated—and that the number of bends in a cube figure, a presentation in perspective, and the presence of bindings of cube figures in different directions should be considered complexity increasing factors that allow for systematic variation in the creation of stimuli even for young participants. Considering that, we argue that our findings contradict the assumption that the mental rotation of three-dimensional objects is generally too complex for younger children to process successfully (Hoyek et al., 2012; Jansen et al., 2013), but rather that complexity of three-dimensional objects can be reduced adequately and in a systematic way to make their mental rotation suitable for young children (e.g., Hawes et al., 2015).

One of our main interest was to contribute to the still rather unanswered question of potential gender differences in mental rotation skills in primary school children (Bott et al., 2023; Grüßing, 2012; McGee, 1979). In general, our results showed that girls solved a smaller number of tasks in the given time than boys, regardless of their age. This finding is in line with the results by Geisler et al. (2008) who demonstrated gender differences in the choice of strategy in eleven-year-olds: It is plausible to assume that different strategies may indeed vary in processing time, leading to variation in the number of solved items in a given time frame. To follow up on this, research with the MERRO could focus on qualitative (stimulated recall) interview studies asking male and female students about their procedure that led to their answer, or on quantitative eye-tracking studies to gain insights on systematic variation of focus points of girls and boys in particular, or high- and low-performing students in general.

We argue that *complexity* of the stimuli may be foundational in understanding gender differences in mental rotation skills in particular, and spatial cognition in general. With regard to the impact of *difficulty* generating factors, we found no valid indication in our data that underpins the assumption that there is an interaction between gender and age, or gender and rotation angle. Yet, our data suggests that gender effects in mental rotation skills may be most prominent in items with *complexity* generating factors presenting three-dimensional cube figures that were tilted in more than one direction (i.e., the commonly used item type in the MRT; Peters et al., 1995) did show notable differences in difficulty between boys and girls. This could also explain why Rahe and Quaiser Pohl (2019) did not find significant gender differences for the analogues ball items compared to the MRT items: While the three-dimensional view of a cube changes when rotating in space, this does not apply to the plastic appearance of spheres when rotating in space. These stimuli therefore may differ in complexity. This, again, opens up new questions regarding the nature of that difference, whether it can successfully be targeted with specific short-term interventions, and when these differences develop. For these questions, the MERRO may yield a valid and efficient instrument to empirically investigate those aspects. Here, the ability to closely examine the relationship between mental rotation skills and, e.g., mathematical achievement with children in primary school allows for new and more advanced RQs, which may lead to a more nuanced understandings of gender disparities—and, ultimately, an answer to the question *at what age* they occur.

The representation of three-dimensional objects in two-dimensional space combined with the question of plane axis mirroring or plane rotation is associated with a considerable limitation of the MERRO test, which should be discussed. Consider non-twisted stimuli representing cube figures in perspective (stimuli 401-404 and stimuli 501-504; **Figure 2**). Here, three-dimensional rotation systematically lead students to an answer considered incorrect in the understanding of the MERRO test. This is not the case for twisted stimuli (stimuli 405, 406, 505, and 506; **Figure 2**). As the MERRO aims at assessing mental rotation skills in children of a young age where two-dimensional rotation can be considered more plausible than three-dimensional rotation, this potential threat to the validity may be of no severe relevance; in fact, we found no indication in the data that those items would not behave as expected (e.g., systematic answer patterns with wrong answers in the non-twisted items, but not in the twisted items). Future studies could utilize person oriented statistical approaches like profile or cluster analyses to investigate whether (small) subgroups of students exist, that demonstrate such patterns. The mental rotation skills of such potential students would be underestimated by utilizing the MERRO test—which is why we argue for using the MERRO test only for assessing primary school children.

In addition, while this study investigates gender differences in mental rotation ability among primary school children, we acknowledge that our analysis does not explicitly account for potential socio-cultural and educational influences on these disparities. Prior research suggests that factors such as early spatial experiences (e.g., play with construction toys and puzzle-solving activities), classroom instructional practices, and broader societal expectations may contribute to the development of spatial abilities. Since our study does not include measures of these external influences, we cannot determine the extent to which they contribute to the observed patterns. Future research should explore these contextual factors by integrating background variables such as parental support for spatial activities, differences in curriculum exposure, or cultural attitudes toward gender and spatial reasoning. A more comprehensive understanding of these influences could further clarify the developmental trajectory of mental rotation skills in early childhood.

CONCLUSION

The timing of the emergence of gender differences in both spatial abilities and mathematical achievement is an intriguing area for further research. Identifying when (and why) such differences manifest could inform the development of more effective educational interventions, aiming to mitigate gender disparities in critical cognitive skills from an early age. This would not only contribute to a more equitable educational landscape but also enhance our understanding of cognitive development in relation to mental rotation, spatial reasoning, and mathematical proficiency. Our proposed MERRO test may serve as a first supplement to the present instruments suitable for children of younger ages and open directions for further development of instruments

Author contributions: **DR:** conceptualization, investigation, methodology, project administration, resources, validation, visualization, writing–original draft, & writing–review & editing & **FR:** conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, & writing–review & editing. Both authors agreed with the results and conclusions.

Funding: The open access publication was funded by the Publication Fund of the University of Education Freiburg.

Acknowledgments: The authors would like to thank the student research assistants who worked on the project.

Ethical statement: The authors stated that the study does not require any specific ethics committee approval according to national standards. The local school authority, Regional Council of Freiburg, evaluated and approved this study with approval number 7-6499.2. The authors further stated that detailed information about the study's aims and procedures was provided to the parents. All children gave verbal informed consent to take part in the study, after their parents gave written informed consent. Consent forms were signed before participation, ensuring that participants' rights and privacy were fully respected throughout the research process.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

- Adaboh, S., Akpalu, R., Owusu-Darko, I., & Boateng, S. S. (2017). Using courseware instruction to improve junior high school students' spatial visualization skills. *International Electronic Journal of Mathematics Education*, 12(3), 353-365. <https://doi.org/10.29333/iejme/617>
- Barke, H.-D. (1980). Raumvorstellung im naturwissenschaftlichen Unterricht [Spatial conception in science lessons]. *MNU*, 33, Article 129.
- Bartlett, K. A., & Camba, J. D. (2023). Gender differences in spatial ability: A critical review. *Educational Psychology Review*, 35(1), Article 8. <https://doi.org/10.1007/s10648-023-09728-2>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, 21(1), 47-60. <https://doi.org/10.2307/749456>
- Besuden, H. (1999). Raumvorstellung und Geometrieverständnis [Spatial perception and understanding of geometry]. *Mathematische Unterrichtspraxis*, 20(3), 1-10.
- Bodner, G. M., & Guay, R. B. (1997). The Purdue visualization of rotations test. *The Chemical Educator*, 2(4), 1-17. <https://doi.org/10.1007/s00897970138a>
- Bott, H., Poltz, N., & Ehlert, A. (2023). Erfassung mentaler Rotationsleistungen im Grundschulalter: Ein computerbasiertes Testverfahren für die erste bis dritte Klasse (cMR) [Assessment of mental rotation performance in primary school children: A computer-based test for first to third grades (cMR)]. *Diagnostica*, 69(3), 121-132. <https://doi.org/10.1026/0012-1924/a000309>
- Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler intelligence scale for children–Fifth edition. *Personality and Individual Differences*, 86, 1-5. <https://doi.org/10.1016/j.paid.2015.05.020>
- Cheng, Y.-L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, 15(1), 2-11. <https://doi.org/10.1080/15248372.2012.725186>
- Cochran, K. F., & Wheatley, G. H. (1989). Ability and sex-related differences in cognitive strategies on spatial tasks. *The Journal of General Psychology*, 116(1), 43-55. <https://doi.org/10.1080/00221309.1989.9711109>
- Corballis, M. C. (1997). Mental rotation and the right hemisphere. *Brain and Language*, 57(1), 100-121. <https://doi.org/10.1006/brln.1997.1835>
- Ebert, W. M., Jost, L., & Jansen, P. (2024). Gender stereotypes in preschoolers' mental rotation. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1284314>
- Franke, M., & Reinhold, S. (2016). *Didaktik der Geometrie in der Grundschule* [Didactics of geometry in primary school] (3rd Ed.). Springer. <https://doi.org/10.1007/978-3-662-47266-8>
- Freudenthal, H. (1971). Geometry between the devil and the deep sea. *Educational Studies in Mathematics*, 3(3-4), 413-435. <https://doi.org/10.1007/BF00302305>
- Frick, A., Hansen, M. A., & Newcombe, N. S. (2013). Development of mental rotation in 3- to 5-year-old children. *Cognitive Development*, 28(4), 386-399. <https://doi.org/10.1016/j.cogdev.2013.06.002>
- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. BasicBooks.

- Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, 36(6), 556-563. <https://doi.org/10.1016/j.intell.2007.12.003>
- Gilligan-Lee, K. A., Hawes, Z. C. K., & Mix, K. S. (2022). Spatial thinking as the missing piece in mathematics curricula. *npj Science of Learning*, 7, Article 10. <https://doi.org/10.1038/s41539-022-00128-9>
- Grüßing, M. (2012). *Räumliche Fähigkeiten und Mathematikleistung: Eine empirische Studie mit Kindern im 4. Schuljahr* [Spatial abilities and mathematics performance: An empirical study with children in the 4th grade]. Waxmann.
- Guay, R. B., & McDaniel, E. D. (1977). The relationship between mathematics achievement and spatial abilities among elementary school children. *Journal for Research in Mathematics Education*, 8(3), 211-215. <https://doi.org/10.2307/748522>
- 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>
- Hawes, Z., Gilligan-Lee, K. A., & Mix, K. S. (2022). Effects of spatial training on mathematics performance: A meta-analysis. *Developmental Psychology*, 58(1), 112-137. <https://doi.org/10.1037/dev0001281>
- Hawes, Z., LeFevre, J., Xu, C., & Bruce, C. D. (2015). Mental rotation with tangible three-dimensional objects: A new measure sensitive to developmental differences in 4- to 8-year-old children. *Mind, Brain, and Education*, 9(1), 10-18. <https://doi.org/10.1111/mbe.12051>
- Hawes, Z., Moss, J., Caswell, B., Naqvi, S., & MacKinnon, S. (2017). Enhancing children's spatial and numerical skills through a dynamic spatial approach to early geometry instruction: Effects of a 32-week intervention. *Cognition and Instruction*, 35(3), 236-264. <https://doi.org/10.1080/07370008.2017.1323902>
- Hoyek, N., Collet, C., Fargier, P., & Guillot, A. (2012). The use of the Vandenberg and Kuse mental rotation test in children. *Journal of Individual Differences*, 33(1), 62-67. <https://doi.org/10.1027/1614-0001/a000063>
- Jansen, P., Schmelter, A., Quaiser-Pohl, C., Neuburger, S., & Heil, 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>
- Kajda, B., & Kajda, B. M. (2010). *Dyskalkulie und visuell-räumliche Fähigkeiten: Stehen visuell-räumliche Fähigkeiten in einem kausalen Zusammenhang mit Dyskalkulie?* [Dyscalculia and visuospatial abilities: Are visuospatial abilities causally related to dyscalculia?]. Kovač.
- Landerl, K., Vogel, S., & Kaufmann, L. (2022). *Dyskalkulie: Modelle, diagnostik, intervention* [Dyscalculia: Models, diagnostics, intervention] (4th Ed.). Ernst Reinhardt Verlag. <https://doi.org/10.36198/9783838557342>
- Lennon-Maslin, M., & Quaiser-Pohl, C. M. (2024). "It's different for girls!" The role of anxiety, physiological arousal, and subject preferences in primary school children's math and mental rotation performance. *Behavioral Sciences*, 14(9), Article 809. <https://doi.org/10.3390/bs14090809>
- Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Sex differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2), 127-155. <https://doi.org/10.1002/wcs.1380>
- Levine, S. C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex differences in spatial skill. *Developmental Psychology*, 35(4), 940-949. <https://doi.org/10.1037/0012-1649.35.4.940>
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56(6), Article 1479. <https://doi.org/10.2307/1130467>
- Lorenz, J. H. (1992). *Anschaung und Veranschaulichungsmittel im Mathematikunterricht* [Visual aids and illustrations in mathematics lessons]. Hogrefe.
- Maier, P. H. (1999). *Räumliches Vorstellungsvermögen: Ein theoretischer Abriss des Phänomens räumliches Vorstellungsvermögen* [Spatial imagination: A theoretical outline of the phenomenon of spatial imagination] (1st Ed.). Auer.
- Marmor, G. S. (1975). Development of kinetic images: When does the child first represent movement in mental images? *Cognitive Psychology*, 7(4), 548-559. [https://doi.org/10.1016/0010-0285\(75\)90022-5](https://doi.org/10.1016/0010-0285(75)90022-5)
- 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>
- Mix, K. S., Levine, S. C., Cheng, Y.-L., Young, C., Hambrick, D. Z., Ping, R., & Konstantopoulos, S. (2016). Separate but correlated: The latent structure of space and mathematics across development. *Journal of Experimental Psychology: General*, 145(9), 1206-1227. <https://doi.org/10.1037/xge0000182>
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133-142. <https://doi.org/10.1111/j.2041-210x.2012.00261.x>
- Nes, F. V., & Doorman, M. (2011). Fostering young children's spatial structuring ability. *International Electronic Journal of Mathematics Education*, 6(1), 27-39. <https://doi.org/10.29333/iejme/259>
- 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>
- Peters, M., Lehmann, W., Takahira, S., Takeuchi, Y., & Jordan, K. (2006). Mental rotation test performance in four cross-cultural samples (N = 3367): Overall sex differences and the role of academic program in performance. *Cortex*, 42(7), 1005-1014. [https://doi.org/10.1016/S0010-9452\(08\)70206-5](https://doi.org/10.1016/S0010-9452(08)70206-5)

- Plath, M. (2014). *Räumliches Vorstellungsvermögen im vierten Schuljahr: Eine Interviewstudie zu Lösungsstrategien und möglichen Einflussbedingungen auf den Strategieeinsatz* [Spatial imagination in fourth grade: An interview study on solution strategies and possible influencing factors on the use of strategies]. Franzbecker.
- R Core Team. (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. <https://www.R-project.org/>
- Rahe, M., & Quaiser-Pohl, C. (2019). Cubes or pellets in mental-rotation tests: Effects on gender differences and on the performance in a subsequent math test. *Behavioral Sciences, 10*(1), Article 12. <https://doi.org/10.3390/bs10010012>
- Rahe, M., Ruthsatz, V., & Quaiser-Pohl, C. (2021). Influence of the stimulus material on gender differences in a mental-rotation test. *Psychological Research, 85*(8), 2892-2899. <https://doi.org/10.1007/s00426-020-01450-w>
- Reilly, D., Neumann, D. L., & Andrews, G. (2017). Gender differences in spatial ability: Implications for STEM education and approaches to reducing the gender gap for parents and educators. In M. S. Khine (Ed.), *Visual-spatial ability in STEM education* (pp. 195-224). Springer. https://doi.org/10.1007/978-3-319-44385-0_10
- Reinhold, F., Hofer, S., Berkowitz, M., Strohmaier, A., Scheuerer, S., Loch, F., Vogel-Heuser, B., & Reiss, K. (2020). The role of spatial, verbal, numerical, and general reasoning abilities in complex word problem solving for young female and male adults. *Mathematics Education Research Journal, 32*(2), 189-211. <https://doi.org/10.1007/s13394-020-00331-0>
- Resnick, I., Harris, D., Logan, T., & Lowrie, T. (2020). The relation between mathematics achievement and spatial reasoning. *Mathematics Education Research Journal, 32*(2), 171-174. <https://doi.org/10.1007/s13394-020-00338-7>
- Ruthsatz, V., Neuburger, S., Jansen, P., & Quaiser-Pohl, C. (2015). Cars or dolls? Influence of the stereotyped nature of the items on children's mental-rotation performance. *Learning and Individual Differences, 43*, 75-82. <https://doi.org/10.1016/j.lindif.2015.08.016>
- Ruthsatz, V., Rahe, M., Schürmann, L., & Quaiser-Pohl, C. (2019). Girls' Stuff, boys' stuff and mental rotation: Fourth graders rotate faster with gender-congruent stimuli. *Journal of Cognitive Psychology, 31*(2), 225-239. <https://doi.org/10.1080/20445911.2019.1567518>
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701-703. <https://doi.org/10.1126/science.171.3972.701>
- Smith, I. M. (1964). *Spatial ability*. R. R. Knapp.
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174-215. <https://doi.org/10.1037/0278-7393.6.2.174>
- Tam, Y. P., Wong, T. T.-Y., & Chan, W. W. L. (2019). The relation between spatial skills and mathematical abilities: The mediating role of mental number line representation. *Contemporary Educational Psychology, 56*, 14-24. <https://doi.org/10.1016/j.cedpsych.2018.10.007>
- Thurstone, L. L. (1950). Some primary abilities in visual thinking. *Proceedings of the American Philosophical Society, 94*(6), 517-521.
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352-402. <https://doi.org/10.1037/a0028446>
- 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>
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2017). *Link between spatial and mathematical skills across the preschool years*. Wiley-Blackwell.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250-270. <https://doi.org/10.1037/0033-2909.117.2.250>
- Weber, A. M., Bobrowicz, K., Greiff, S., & Leuchter, M. (2024). Mental rotation is supported by block play in boys and girls. *Journal of Applied Developmental Psychology, 91*, Article 101630. <https://doi.org/10.1016/j.appdev.2023.101630>
- Xie, F., Zhang, L., Chen, X., & Xin, Z. (2020). Is spatial ability related to mathematical ability: A meta-analysis. *Educational Psychology Review, 32*(1), 113-155. <https://doi.org/10.1007/s10648-019-09496-y>