Mental Rotation

The Test Design, Sex Differences, and the Link to Physical Activity

Inaugural-Dissertation zur Erlangung der Doktorwürde der

Fakultät für Humanwissenschaften der Universität Regensburg,

vorgelegt von

Leonardo Jost

aus Bochum

-2021-

Regensburg, 2021

Gutachterin (Betreuerin): Prof. Dr. Petra Jansen

Gutachterin: Prof. Dr. Claudia Quaiser-Pohl

Content

Acknowledgements

As much as we deal with randomness of data in statistics, there likely is the same randomness in life. Despite not knowing the outcomes of alternatives, I must acknowledge how lucky I have been. Through the series of improbable events that thus is life, I have been granted the chances and opportunities that have ultimately led me to write (and complete) this thesis.

Most importantly, about three years ago, Petra Jansen allowed me to start working on this thesis and continuously supported me. I am very thankful for her guidance and faith in my ideas for the experiments presented here and for all the helpful comments that shaped them into presentable versions and allowed me to learn so much in this time.

The foundations that allowed me to take these chances were of course laid much earlier. Be it through nature or nurture, but also their constant care and support in any way possible, I want to thank my parents for all the possibilities they have opened for me throughout my life. Along the way, there were an uncountable further number of family, friends, colleagues, teachers, and many more, who supported me and from whom I learned many life lessons. While much of that knowledge was not directly applied here, they were invaluable in shaping me into the person I am today and without them, I would not be here.

Summary

Spatial abilities support us in navigating and understanding the space that our world consists of. They are educationally relevant through their specific link to mathematical performance and STEM attainment (Xie et al., 2020). One spatial ability is the ability to imagine objects or images in different orientations in the mind. Interestingly, it seems to follow the same time-path as physical rotations with a constant rotation speed and it is accordingly named mental rotation (R. N. Shepard & Metzler, 1971). There are many open questions regarding it, including the interaction with physical activity, sex differences in performance, but also how to best measure this ability. These are explored in this thesis.

In the first study, we investigated the performance during of simultaneous mental rotation and aerobic exercise. Compared with isolated mental rotation and isolated aerobic exercise, we found higher subjective cognitive and higher objective physiological effort, respectively. These are in support of neurological models describing conflicts between differing demands of implicit and explicit tasks (Dietrich & Audiffren, 2011).

In the second study, we further investigated the chronometric mental rotation test employed in the first study. Because mirrored stimuli cannot be rotated into congruence, they are typically discarded from analysis (R. N. Shepard & Metzler, 1971). We circumvented this drawback by presenting two alternatives such that congruence by rotation can be achieved in all trials. The results suggest that indeed mental rotation processes are used in all trials. The design can thus improve the analysis of accuracies while increasing power.

The first two studies both showed an improvement over time during the sessions themselves. Similarly to differing pretest performances, differing improvements within tests between different groups can impact the detection of improvements between tests due to a treatment. In the third study, we attempted to better separate these effects and investigated the

statistical relevance by simulations on the dataset of the second study. The results demonstrate both the impact of within-session practice effects and the usefulness of the presented approach.

The design of the second study and the analysis of the third study were applied in the fourth study for the investigation of manual training of mental rotation. Manual rotation interventions have been proven to improve mental rotation performance but always employ a concurrent visual rotation (Adams et al., 2014; Wiedenbauer et al., 2007). By separating the congruent and causal motor activity and the visual rotation, we created three interventions. The comparison showed a similar improvement for all groups. Due to the independence of the form and occurrence of the manual activity, this suggests that it is not the motor activity but the concurrent visual rotation that leads to improvements in mental rotation tasks.

Sex differences in mental rotation performance are a topic of great interest due to the link to STEM attainment but the reasons are not clear. In the first studies, we reported some results regarding sex differences but did not investigate the underlying reasons. One suspected reason is the stimulus material (Rahe et al., 2020), which we turned to in the fifth study. As emotions are also known to influence test performance, we investigated the implicit affective evaluation of the stimuli. For the cube figures but not for the pellet figures, the mental rotation performance was significantly predicted by the implicit affective evaluation of the respective figures. Implicit attitudes could thus be a reason for varying sex differences between stimuli but need further investigation.

The large sex differences in mental rotation are, however, mostly found in the psychometric test and not in the chronometric test (Peters & Battista, 2008). As we also did not find performance differences in the fourth study for our design of the second study, we compared the designs in detail in the sixth study. The differences regarding the number of alternatives and whether the alternatives were pairwise mirrored were further experimentally investigated for their effect on sex differences. At least for the participants of main interest, the results did reveal such an effect: Sex differences were smaller when the alternatives were presented as pairwise mirrored. This indicates

that sex differences are indeed at least in part due to the test design and cannot be attributed solely to mental rotation ability, which is credibly involved in all tests. The combination of the test design, the link to manual rotation, as well as neurological resources can help us better understand mental rotation ability and its' phases. They could allow us to identify other abilities involved in solving mental rotation tests, separate these abilities, and pinpoint the reasons for sex differences.

1 Preface

Spatial abilities refer to skills regarding the mental representation and transformation of objects and shapes and their relationships. They are an interesting research topic not only because of their usefulness in everyday life but also for their relationship to general intelligence and the predictability of academic and professional success in STEM-related areas by spatial ability performance. One spatial ability is mental rotation, the ability to imagine the appearance of rotated objects, which has been studied extensively since its scientific inception by R.N. Shepard and Metzler in 1971. Originally studied as the imaginary rotation of abstract three-dimensional objects it has since been applied to many different both abstract and realistic, two- and three-dimensional objects and results robustly indicate a similarity to physical rotations at fixed rotation speeds. Despite its scientific popularity for both theoretic insights and practical applications, there are of course still many open questions.

One question concerns the measurement of mental rotation ability, for which two widely used tests exist. While each has advantages and disadvantages due to the construction for different main application areas and analyses, by design both plausibly measure the ability to rotate and compare objects in the mind. Interestingly however, there is a difference in the male performance advantage between tests. Next to the topic of sex differences in mental rotation performance, this indicates open questions and possible improvements about the measurement of mental rotation ability.

Due to the relation to STEM fields, which throughout time have shown larger male involvement, sex differences in spatial abilities are an important topic to promote equality. For mental rotation, many factors have been investigated and identified as possible reasons. These include both biological factors related to sex as well as social factors related to gender. However, due to varying differences between mental rotation tests, the reasons for performance differences between sexes are obviously not yet fully understood.

Another interesting topic about mental rotation is the relation to motor abilities. The similarities to physical rotation have been investigated and congruent rotational hand movements have shown very similar patterns. Training rotational hand movements has also been shown to enhance mental rotation performance. These insights offer the possibility to further investigate the processes of mental and physical rotation and further analyzing the effects of rotational hand movements could enhance our understanding of the underlying processes behind mental rotation. While mental rotation has a special relationship to physical rotation due to this similarity, a general question of interest is how cognition and overall physical activity interfere or can enhance each other. It is often postulated that physical activity can improve our cognitive functioning but there are still many questions regarding the exact mechanisms. Studying the interaction of mental rotation and general physical ability can enhance our comprehension of the overall relationship between cognition and physical activity but also the general classification of mental rotation.

As demonstrated, these three fields of the test design, sex differences in performance, and the relationship to physical rotation (and physical activity) are connected. This thesis attempts to contribute to our understanding of mental rotation ability by investigating these topics.

2 Theoretical Background and State of Research

This chapter gives a short introduction into spatial abilities in general and mental rotation in particular. It will introduce the prevalent tests and differences between them as well as possible influence factors regarding results. Furthermore, the connection to motor abilities is described as the basis for applications.

2.1 Spatial Abilities

Spatial abilities are generally viewed as an important component of intelligence as well as an important part of functioning in everyday life, given that our world exists in space (Buckley et al., 2018; Newcombe & Shipley, 2015). For example, the estimation and recognition of distances and objects are important for the orientation in ones surrounding and the use of tools. However, before delving into the details of spatial abilities and mental rotation, it is important to know what spatial abilities are and how different parts of spatial abilities can be separated.

Regarding the classification of spatial abilities, psychometric approaches have tried to characterize spatial abilities into the three parts of spatial perception, spatial visualization, and a separated category for mental rotation (Linn & Petersen, 1985). Other works have also introduced more possible (sub-)categories such as spatial orientation or spatial relations and different names for the same or similar categories. However, the definition of and distinction between these groups is not entirely clear (Hegarty & Waller, 2005; Kozhevnikov & Hegarty, 2001; Xie et al., 2020). Thus, instead of a psychometric classification, Uttal et al. (2013) and Newcombe and Shipley (2015) proposed to distinguish between fundamental task characteristics and identified two factors. The first factor differentiates between intrinsic and extrinsic information. Intrinsic spatial information describes the relation and spatial orientation of an object itself and its parts, whereas extrinsic information describes an object in relation to other objects. The second factor differentiates between static and dynamic tasks, that is, whether the object or objects in question are moved and change their position or orientation.

By this classification, intrinsic dynamic tasks require a mental transformation of an object itself, such as a rotation for the mental rotation test (R. N. Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). The complexity of intrinsic static tasks on the other hand lies in the perception of objects amidst distracting information. One example is the identification of simple figures embedded into more complex figures in the embedded figures task (Witkin, 1971). Extrinsic static tasks require the understanding of abstract spatial principles independent of the object in question such as the horizontal invariance of the water level in a tilted bottle in Piaget's water level task (Piaget & Inhelder, 1956). In extrinsic dynamic tasks, participants have to imagine an environment from a different perspective as for example in the Guilford-Zimmerman spatial orientation test (Guilford & Zimmerman, 1948). In this test participants must identify the position of a boat such that a given view of the landscape is achieved.

The proposed 2x2 classification is of course not exempt from scrutiny. Mix et al. (2018) questioned the static-dynamic distinction as it did not prove to be a significant factor in a confirmatory factor analysis in three samples of spatial performance of kindergarten, third grade, and sixth grade students. A review of Buckley et al. (2018) identified the additional category of visual-spatial working memory but also addressed the need for further investigation and distinction especially within and between static and dynamic spatial abilities.

Notwithstanding the yet unclear classification, spatial abilities are already widely studied for their practical implications. Most importantly, it has been shown that spatial abilities are an important predictor of achievement and attainment in STEM (science, technology, engineering, and mathematics) domains (Buckley et al., 2018; Shea et al., 2001; Wai et al., 2009, 2010) and may be a reason for the male advantage in arithmetical reasoning (Geary et al., 2000). Xie et al. (2020) further demonstrated that the relationship to mathematical abilities is not limited to specific spatial abilities but holds for all categories defined by Uttal et al. (2013) and the additional category of visual-spatial memory. While the exact mechanisms of the relationship between factors of spatial abilities and STEM abilities are not yet uncovered due to the complexity and diversity of both, the training of spatial abilities has emerged as a tool to improve STEM performance. Due to the malleability of spatial skills and the transfer of training effects of one spatial ability not only to spatial tasks in the same class but also in general, a spatial training can also promote STEM success (Uttal et al., 2013). Thus, despite some uncertainties regarding the classification of spatial abilities, the study of both individual spatial abilities as well as general spatial performance is of scientific interest.

2.2 Mental Rotation Ability

Mental rotation describes the cognitive ability to rotate objects or images in the mind, which by the classification of Uttal et al. (2013) is an intrinsic dynamic spatial ability. Buckley et al. (2018) however differentiate between further factors and argue that mental rotation tests can tap into different factors and this might also differ between different tests of mental rotation ability. One of these factors is perspective taking. While closely related to mental rotation factors (Hegarty & Waller, 2004), perspective taking is a separate factor and would also fall into the extrinsic dynamic category of Uttal et al. (2013). Despite these factors being different, they might not be fully separable in tests as participants are observed to use strategies of both mental rotation and perspective taking to solve mental rotation tasks (Hegarty, 2018; Hegarty & Waller, 2004). Another necessary factor for mental rotation tasks is the visual working memory as a storage of object representations. Interestingly, Hyun and Luck (2007) identified the object working memory subsystem instead of the spatial working memory subsystem being primarily used as information about the objects themselves is stored and not about the operations performed on them. Mental rotation tests are also not limited to the ability of mental rotation itself but typically consist of an identification or discrimination task based on the rotation and thus further include multiple functionally independent processing stages (Heil & Rolke, 2002; R. N. Shepard & Cooper, 1986). Nevertheless, it is undeniable that mental rotation is a spatial ability and as with all spatial abilities, the performance in mental rotation tests shows strong links to STEM performance (e.g. Hausmann, 2014; Moè, 2016; Moè et al., 2018).

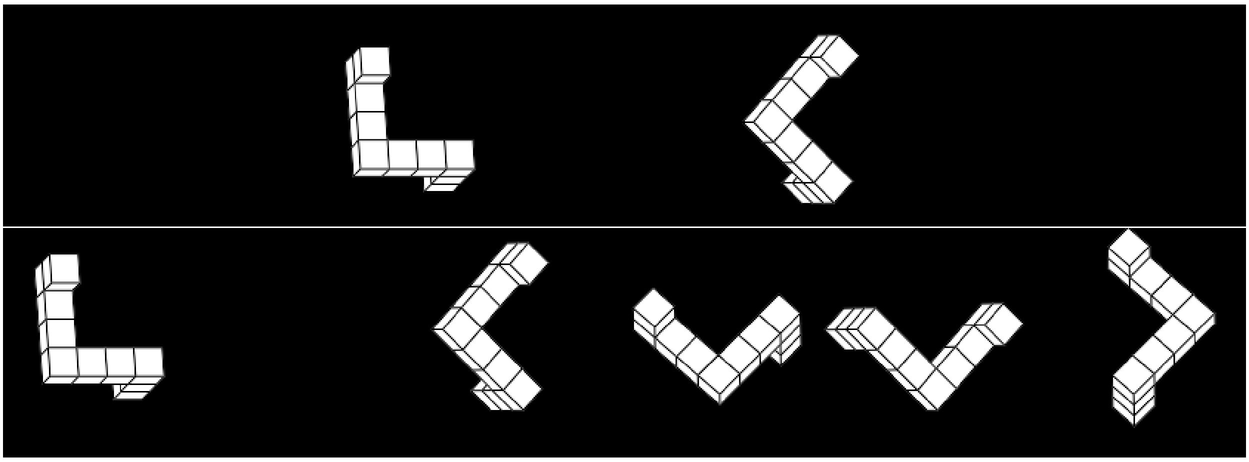
Zacks et al. (2000) describe two different classes of mental rotary transformations: object-based and egocentric transformations. In object-based transformations, the observer's position remains fixed, and the object is rotated mentally in relation to the surrounding environment. In egocentric transformations, the observer is tasked to change their own perspective by an imaginary rotation of their own body. Both transformations are typically represented by different tests. Whereas object-based mental rotation tests typically require the comparison of figures in different rotary states, egocentric tests require a left-right judgement of a single object with a clear upright orientation, from which it differs by a rotation. The distinction, which kind of transformations is actually used by participants, is however also dependent on instructions and stimulus material (Voyer, Jansen, et al., 2017; Zacks et al., 2000) and is also not uniform between participants and trials as evidenced from the use of multiple strategies (Hegarty, 2018; Hegarty & Waller, 2004). While both test forms emerged from imaginary rotations and egocentric tests are often labeled as egocentric mental rotation tests, they fall into different categories of spatial abilities by both the classifications of Uttal et al. (2013) and Buckley et al. (2018). This difference is also evident in the difference of results revealed by both tests (Voyer, Jansen, et al., 2017; Zacks et al., 2000). In the following, we will thus focus on mental rotation tests employing object-based transformations.

2.3 Mental Rotation Tests

Two types of mental rotation tests are widely used to assess mental rotation performance: 1) chronometric tests (figure 1, top) based on the study of R. N. Shepard and Metzler (1971), where two objects are presented which are either same (rotated) or different (mirrored). 2) psychometric tests (figure 1, bottom) based on the study of Vandenberg and Kuse (1978), where compared to one target, two out of four alternatives are rotated and the other two are mirrored or structurally different.

Figure 1

Examples of Trials of Chronometric (Top) and Psychometric (Bottom) Mental Rotation Tests.



2.3.1 Chronometric Tests

In each task of chronometric mental rotation tests, participants are shown two objects and must decide, whether the objects are the same (i.e., one can be transformed into the other by a rotation) or different (i.e., mirrored and neither can be transformed into the other by a rotation). The chronometric test is especially interesting because it allows the analysis of the rotation process of individual items as for every task the reaction time and correctness of the answer is recorded. In their seminal study, R. N. Shepard and Metzler (1971) identified a linear relationship of reaction time and the angular disparity of the two objects, which has since then been replicated extensively although not always as perfectly linear. This proportionality indicates that the objects are indeed rotated mentally in a continuous way with the slope representing a somewhat constant rotation speed. The intercept of this linear relationship represents the time needed for nonrotational processing stages to identify two identical images (Heil & Rolke, 2002). Similar to but less pronounced than the relationship of angular disparity and reaction time, error rates also increase monotonously and roughly linearly with increasing rotation angles. While there is no straightforward explanation as the link to rotation speed, this effect ensures that slower reaction times are not due to speed-accuracy trade-offs. One possible explanation is the neural interference of the mental representation of the rotation with the storing of the representation in the working memory (Carpenter et al., 1999; Hyun & Luck, 2007).

Mental rotation tests have been extensively studied with both two-dimensional and three-dimensional stimuli, both rendered as two-dimensional images. The most popular stimuli are three-dimensional cube figures similar to the ones employed in the original study of R. N. Shepard and Metzler (1971), which consist of 10 connected cubes. These figures have been reused and recreated by many researchers. In an attempt to further standardize the stimulus material and provide further options, Peters and Battista (2008) created an extensive library of cube figures, which since then has been widely used. The library consists of 16 different figures rotated around all canonical axes in steps of 5° for two orientations (basic and mirrored or a and b orientation in the library), in two different colorings (only white cubes or alternating white and grey cubes), and with two different backgrounds (white and grey/black) for a total of 28032 images.

As with most reaction time measurements in cognitive tests, the reaction time in chronometric tests is usually only analyzed for correct answers because for wrong answers it is unknown why an error occurred. To avoid having to deal with missing data in the statistical analysis using ANOVAs, chronometric tests are typically conducted using a given and typically large number of stimulus pairs, which are repeated until each pair is answered correctly once by every participant. Due to the need for computers to conduct the tests, participants are most often tested individually, and early studies employed only small numbers of subjects.

There is one main methodological concern in the use of chronometric mental rotation tests: Because mirrored figures cannot be brought into congruence those items have been deleted mostly from analysis so far (Jolicœur et al., 1985; R. N. Shepard & Metzler, 1971). This incongruence is rarely further analyzed but is reflected in reaction time differences of almost 1s in the analysis of R.N. Shepard and Metzler (1971) and around 500ms for cube figures in the study of Voyer and Jansen (2016). As mirrored figures typically contribute about half of the tasks, disregarding them means losing a noticeable amount of power or testing time in experiments (Brysbaert & Stevens, 2018). Moreover, analyzing accuracy only on non-mirrored stimuli disregards both false alarms and correct rejects, which might show a different pattern than hits and misses. Analyzing sensitivity instead of accuracy incorporates false alarms (Stanislaw & Todorov, 1999). However, as false alarms occur only on mirrored stimuli, they cannot be matched to an angle of rotation and only global analyses of sensitivity are appropriate.

2.3.2 Psychometric Tests

In psychometric mental rotation tests, participants are shown five figures in every task. One figure, typically to the left of the other figures, is the target figure. Out of the four other figures, named alternatives or comparison figures, exactly two are rotated to the target figure and the other two are different. The different alternatives are either mirrored or structurally different to the target figure. Participants are then tasked to identify the two rotated alternatives in every trial. The psychometric tests were originally developed as paper and pencil tests by Vandenberg and Kuse (1978) using similar cube figures as the chronometric test. The main advantage of the psychometric test was the possibility to quickly test large groups of subjects.

As with the chronometric tests, many researchers have copied and reused the original test version. Due to the deterioration of the test material through successive physical copying, Peters et al. (1995) created a redrawn version. Two of these test versions, the A and B version consisting of the same trials but arranged in different orders. Peters et al. also provide a third version (C), which employs rotations around two axes and is thus much more complex. In these most used versions, the test consists of two sets of twelve tasks preceded by three practice trials. Each set of twelve tasks must be completed within three minutes with unanswered trials typically considered as wrongly answered. To reduce the impact of guessing items correctly, the test awards one point per trial only if both correct alternatives are selected.

2.4 Sex Differences in Mental Rotation Test Performance

One of the most interesting observation about psychometric tests is one of the largest sex differences in performance known in cognitive psychology (Voyer et al., 1995). These are often investigated to understand differences in spatial abilities and their transfer to STEM performance, but the exact causes are not known. In contrast to psychometric test performance, differences between sexes in chronometric tests are much smaller and mostly non-significant or only observed on certain subtests or specific stimuli (Jansen-Osmann & Heil, 2007b; Peters & Battista, 2008). While the performance of different mental rotation tests are correlated (Voyer et al., 2006), the reasons for the conflicting results concerning possible sex differences are not completely understood.

Research has focused on examining gender and/or sex differences in VK tests relating to biological factors such as hormones, menstrual cycle, and sexual orientation (Hausmann et al., 2000, 2009; Peters et al., 2007; Peters, Laeng, et al., 1995), social factors such as gender stereotypes and stimulus familiarity (Hausmann et al., 2009; Ruthsatz et al., 2014, 2015, 2017), education and academic background (Peters et al., 2007; Peters, Laeng, et al., 1995), or strategy selection (Hegarty, 2018; Heil & Jansen-Osmann, 2008; Voyer et al., 2020). There is also the question whether these performance differences are caused by gender or sex. Studies investigating biological factors would indicate sex differences, whereas social factors would suggest differences by gender. However, most participants in studies are cis-gender persons and these might be confounded. The question remains why one test design provokes these factors to manifest in performance differences by sex while the other design does not. Peters and Battista (2008) already note that these differences might not be related to differences in actual mental rotation ability, but for example to the switching between pairwise comparisons within one trial for VK tests or dwelling on details of the stimuli. However, research investigating the test design has been scarcer but has analyzed effects of the existence of a time limit (e.g. Peters, 2005; Voyer, 2011), the answering format (two out of four choice vs. same/different, Titze et al., 2010) or the distractor configuration (Voyer & Hou, 2006).

While a relaxation of time limits and some aspects of the distractor configuration have been found to reduce performance differences between sexes, neither can fully explain the differences between tests.

Many of the aforementioned studies indeed identified interactions with sex differences. However, if these factors influenced the mental rotation process, which both tests are supposed to measure, the same performance difference between sexes should occur in chronometric tests as most of those factors are also applicable. Nevertheless, further investigating sex differences in psychometric test performance is interesting as these are some of the largest in cognitive psychology (Halpern, 2012). Despite them not being clearly related to the mental rotation process, our understanding of these sex differences can enhance our understanding of cognitive sex differences in general. Two of these related to emotional aspects of the tests shall be reviewed further in the following.

2.4.1 Stereotype Threat and Social Factors

Stereotype threat describes the fear of confirming a negative stereotype about a group which one belongs to and has been shown to negatively influence performance (Heil et al., 2012; Steele & Aronson, 1995). A complementing effect is the stereotype lift, which describes a performance boost by negative stereotypes about an outgroup (Walton & Cohen, 2003). These can influence mental rotation performance on multiple levels. The used stimuli can be seen as more male stereotyped and spatial abilities in general can be seen as more male stereotyped. Moreover, STEM interest and performance is also linked to the stereotype of better spatial abilities.

For stereotyped mental rotation objects, for example cars or dolls, a significant interaction between the gender of fourth graders and the type of object (male vs female stereotyped) was detected (Ruthsatz et al., 2017). For the comparison of abstract stereotyped figures, Ruthsatz et al. (2014) developed a stimulus set of pellet figures where each cube of a cube figure was replaced by a round pellet. Compared with the male stereotyped cubed figures these were seen as female stereotyped by 10-year-old children. In the mental rotation performance, there was no significant gender difference for the pellet figures compared with a large difference for the cube figures, which was especially noticeable for the rotations in depth. Rahe and Quaiser-Pohl (2020) and Rahe et al. (2020) partially transferred these results to adult participants. They used only the figures rotated in depth and found large performance differences favoring men for cube figures and possibly smaller differences for pellet figures but no significant interaction.

While for the stimulus material, it is assumed that stereotyped objects lead to better performance for the favored gender, more complex effects have been found when a stereotype was used in the instructions. Moè and Pazzaglia (2006) and Heil et al. (2012) found worse performance when participants were instructed about the superiority of the opposite gender. By further investigating the guessing behavior it could be demonstrated that the performance of the women followed the gender belief induction, but their guessing behavior was not affected. For the men the result was inverted: Their guessing behavior followed the stereotype instruction, whereas their performance remained unaffected (Heil et al., 2012). However, in those studies explicit stereotype investigations were used. There was only one study in which the effect of an implicit intervention has been investigated: Guizzo et al. (2019) showed that in the stereotype-nullifying condition a higher automatic association between space and men was linked to a lower performance of the men.

The effect of stereotype threats is not limited to the test itself but applies to spatial abilities in general. By implicitly priming gender stereotypes about spatial and verbal abilities using a questionnaire, Hausmann et al. (2009) found a large effect of these implicit gender stereotypes for gender differences on mental rotation performance. For the explicit answers on the questionnaire, the perception of a stereotype for spatial abilities correlated with performance, at least for women. However, Hausmann (2014) identified that the confidence in one’s own abilities was a better predictor of mental rotation performance than the perception of a stereotype about the gender one belongs to.

Next to stereotyping another social explanation for the mental rotation performance is the use of spatial toys (Moè et al., 2018). Women in STEM degrees had a better mental rotation performance if they had preferred spatial toys in childhood. This is in line with the study of Voyer et al. (2000), who already demonstrated that both women and men had a better mental rotation performance when they preferred spatial toys in childhood.

2.4.2 Explicit and Implicit Affective Evaluations and Mental Rotation

Next to effects linked to the specific test material, there is also a link of mental rotation performance to emotion in general. As geometric figures have been shown to induce affective evaluations, the effect of implicit emotions could be a link between the effects of the stimulus material and the overall emotional state on performance.

Two studies have investigated the effect of explicitly induced emotions by presenting fearful pictures: Borst et al. (2012) demonstrated that participants with high state anxiety rotated cube figures from a mental rotation test more quickly at a comparable error rate after they saw fearful faces than after they saw neutral faces. This was not the case for participants with low state anxiety. They conclude that fear cannot only improve mental rotation but also that these effects depend on the emotional arousal of the participants. However, in a study of Kaltner and Jansen (2014) this result could be specified: the enhancement of mental rotation performance by fear was limited to egocentric mental rotation tasks. In general, participants with high scores on the trait-anxiety scale showed poor results in both reaction time and mental rotation speed.

The isolated influence of implicitly induced emotions on mental rotation performance has been investigated by Mammarella (2011). Before each task of the psychometric mental rotation test a subliminal priming of a happy, sad, or neutral smiley masked by nonsense geometrical features was presented. The results did not indicate differences between the happy and sad subliminal presented condition but showed a general increase in mental rotation ability after unconscious exposure to emotional stimuli, independent of the valence of the emotion. Mammarella assumed that unconscious emotions influence imagery. But until now, it has not been investigated if not only implicit emotions, but also implicit affective evaluations are related to mental rotation performance. Those evaluations are often investigated with an implicit affective priming paradigm.

Affective priming effects reflect the participants' implicit attitudes toward a primed object or person (De Houwer et al., 2009). There are several explanations for the existence of an affective priming effect: One explanation is that a prime can automatically produce an affective evaluation, which involves a "spreading activation mechanism": If the subsequent target has a congruent valence, the response will be facilitated because the response pathway has already been activated. In contrast, targets with an incongruent valence will initiate the wrong response pathway, which needs to be inhibited first to enable the correct response (Fazio, 2001). However, other explanations assume a connectivity network (Spruyt et al., 2002) or a pre-activation of the target through semantic, as well as a response priming (Eder et al., 2012).

Wang and Zhang (2016) used an affective priming paradigm to investigate the implicit affective attitude towards geometric shapes. Their results provide evidence that downward triangles are perceived as negative and circle as positive. They stated that their emotional meaning can be activated automatically and have an influence on the electrophysical processing of subsequent stimuli. Palumbo et al. (2015) demonstrated that curved polygons were given female names and were associated with positive concepts whereas angular polygons were given male names and were associated with negative concepts. Larson et al. (2012) came to the result that participants judged downward-pointing triangles faster as unpleasant compared to neutral or pleasant. This implies that geometric forms convey emotion.

2.5 Mental Rotation and Motor Abilities

The relationship between mental rotation and motor abilities can be analyzed in two ways. First, there is a general interaction of cognitive and motor abilities and especially aerobic exercise. Second, there is a link of mental rotation to rotational movements. An interesting question for both topics is also if and how mental rotation affects the motor abilities. While the interaction of exercise and cognition has been investigated for both acute and chronic effects, the focus here will be on the immediate interaction of both. For exercise, this means the simultaneous performance of both exercise and cognitive tasks. For rotational movements, this also means immediate training effects next to simultaneous execution of both.

2.5.1 Simultaneous Exercise and Cognitive Performance

Since both cognitive and physical performance place demands on the brain, it seems plausible that they interfere with each other during simultaneous execution. Although not unambiguously, this interference has been found in many studies. Lambourne and Tomporowski (2010) and Chang et al. (2012) summarize diverse impairments and facilitations of acute aerobic exercise on cognitive performance. Dual-task studies of simultaneous exercise and cognitive tasks demonstrate that the stability and accuracy of cognitively demanding motor tasks is affected, for example, balance and gait behavior for walking in older adults. Dual-task costs generally describe impairments under dual-task conditions, although facilitations are sometimes also observed (Riby et al., 2004; Schaefer et al., 2015; Schaefer & Schumacher, 2011; Schäfer et al., 2006).

Based on the allocation of neural resources, the strength model of self-control assumes a limited resource of self-control and predicts a reduced performance for any activity demanding this resource once it is depleted (Audiffren & André, 2015; Baumeister et al., 1998, 2007; Englert, 2016). In line with this prediction, the dual-task cost effect for simultaneous performance of cognitive and motor tasks is greatest for a combination of tasks that must be solved attentively such as coordinative movement tasks and memory tasks and for older individuals for whom even simpler motor tasks are comparatively demanding (Riby et al., 2004; Schäfer et al., 2006). A complementary approach to the lower costs and, in some cases, even positive effects of cognitively easy tasks is provided by the reticular-activating hypofrontality (RAH) model of acute exercise (Dietrich & Audiffren, 2011). This model makes a stronger distinction between explicit tasks that require active attention and implicit tasks that can be solved automatically. Implicit and explicit tasks are controlled by different brain areas, both of which require limited metabolic resources. Thus, stressing one system results in reduced performance of the other. Unlike the possibility of arbitrary prioritization of our attention to a task, the allocation of resources to the two systems occurs involuntarily. Also, in contrast to a necessary division of our attention in the presence of multiple explicit tasks, strengthening the implicit system benefits all implicit tasks.

While the strength model is undirected as to which activity drains self-control and thus would also predict a decline in exercise performance if self-control is depleted after the use of executive functions, the RAH model in theory only analyzes the effect of exercise on cognitive performance. However, as explicit cognitive tasks are associated with brain activity which opposes the exercise facilitating brain activity proposed by the RAH model, a decline in physical performance seems feasible. This partially aligns with the observed experimental results; as cognitive tasks become more complex, they might require more effort from the explicit system in line with larger observed dual-task costs. Similarly, as motor tasks become more cognitively demanding due to age or less demanding due to a higher fitness level, they could demand more or less effort, respectively, from the explicit system. Both models would predict increased dual-task costs in the elderly and lower cognitive performance in less fit individuals, as is typically observed. There is further empirical evidence that our cognition indeed influences physical performance. For example, mental fatigue induced by a cognitive task prior to exercise increases the subjective perception of effort (Martin et al., 2018; Van Cutsem et al., 2017) and physiological parameters including heart rate are affected by cognitive tasks in resting conditions (Kahneman et al., 1969; Kennedy & Scholey, 2000). Moreover, the predictions of the models are also not linked to a physical and simultaneously performed cognitive tasks but can also be applied to two cognitive or two motor tasks. This is consistent with reciprocal impairment when multiple cognitive tasks are performed simultaneously (Riby et al., 2004).

Nevertheless, there exists contradicting evidence for the models, both in the predictions of cognitive outcome measures as in part in the aforementioned meta-analyses (Chang et al., 2012; Lambourne & Tomporowski, 2010), as well as in measurements of brain activity (e.g. Dodwell et al., 2019). For a visual working memory task during moderate aerobic activity on a treadmill and a bicycle, Dodwell et al. found better performance compared to seated and standing control conditions. Using EEG, they also failed to find evidence for neural resource conflicts and even facilitation during processing stages, which should require the explicit system. However, it is possible that the cognitive tasks were too easy, and performance was not hindered by the neural resource reallocations predicted by the RAH. If performance is not limited by neural resources, overall performance might be more influenced by how much effort participants are willing to invest, either consciously or subconsciously. Dodwell et al. further stated that performance facilitation might also occur due to differences in overall arousal, which is not a subject of the RAH. The larger facilitation in the standing and running conditions might also be due to an overlap of neural mechanisms governing balance and spatial cognition. This suggests that neural resource allocation does indeed influence performance but the RAH does not account for these overlaps.

2.5.2 Manual Rotation and Mental Rotation

Because the assumed mental rotation process using a rotation speed resembles the process of physically rotating an object there is a possible link between both. In their classic study, Wohlschläger and Wohlschläger (1998) investigated the similarity between mental and manual object rotations. They did find similar reaction times for both mental and manual rotation around cartesian axes. In a second experiment, an interaction was found between mental object rotations and simultaneously executed hand movements but only for rotational hand movements around the same rotation axis but not for rotational movements around another axis, translational hand movements, or the mere presentation of a directed arrow. The mental rotation was faster if the manual rotation was performed in the shortest direction for the mental rotation and an interference was observed for opposite directions and these effects were larger for the dominant hand. The authors suggested a common process in mental object rotation and the programming of hand movements. In a later study, Wohlschläger (2001) showed that motor planning is a crucial factor whereas preparation and execution are not as critical to performance. He concluded that mental rotation is an imagined (covert) motor action and that the interference he observed in his studies represents interference between incompatible actions. Furthermore, Gardony et al. (2014) found similar angular disparity effects in mental and physical rotations. In their study participants had to decide if two rotated objects on a screen were the same or different either mentally or while rotating a bimanually held sensor. Their analysis demonstrated an increase of reaction time and error rate with increasing angular disparity for the "same" trials but not for "different" trials. Those studies, and also the study of Wexler et al. (1998) who demonstrated that compatible manual and mental rotation results in faster reaction times and fewer errors have triggered studies on the importance of motor processes in mental rotation and suggested a link between motor preparation and mental rotation. However, Janczyk et al. (2012) demonstrated in a series of studies that the connection between manual rotation movements and mental rotation could instead be the anticipated sensory output of the manual rotation.

Due to the similarity between mental and manual rotation, manual rotation has also been investigated as a training for mental rotation. For such a training, Wiedenbauer et al. (2007) developed a manual rotation training in which participants had to rotate a joystick, which was hidden in a box. Using the joystick one of the two cube figures presented on a screen should be rotated into congruence with the other figure. The training included 192 trials, and the control group played a non-spatial computer game. Before and after the training a chronometric mental rotation test had to be completed. The manual rotation training did indeed improve mental rotation performance but specifically for those objects, which were used for training. Adams et al. (2014) replicated these results in a similar experiment. They also showed that this improvement by the manual rotation training was comparable to an improvement by repeated mental rotation tasks, further indicating that manual rotation contains similar aspects as mental rotation. The manual rotation performance, however, only improved through a manual training but not a mental training. This suggests that there are parts of the manual rotation process, which are not trained by mental rotation. The effect of manual training on mental rotation was also observed in children

(Wiedenbauer & Jansen-Osmann, 2008). While these studies indicate a training effect of manual on mental rotation and despite the link between manual and mental rotation, it is not clear that the training effect is specifically caused by the motor activity. All manual rotation trainings incorporate a visualization of the rotation, which is causally linked to the movement. In addition, it is possible that the participants rotated the stimuli mentally to plan the manual rotation, which would explain the training effect on mental rotation performance.

2.6 Method and Analysis

While the previous chapters have discussed interesting results regarding mental rotation specifically, an ongoing problem is the overall confidence in scientific results and there are two important points to discuss. First, results need to be detectable. Second, results need to be replicable. Especially the second point has proven to be problematic in the past. In combination with a publication bias favoring significant results, the use of questionable research practices and wrong interpretations of undetected effects may have led to many exaggerated or even wrong conclusions in past research (e.g. Amrhein et al., 2019; Ferguson & Heene, 2012; Pashler & Wagenmakers, 2012). The task of current and future research can thus not only be to generate interesting results, but it must be to also raise confidence in these results. Although it is not the main topic of this thesis and there is much more depth to these aspects than can be presented here, it is thus important to acknowledge these problems and to implement measures to reduce negative effects.

2.6.1 Detectability

Due to a publication bias in research, the detection of effects is typically not a problem but rather by what means they are detected. However, it is important that the correct effects are detected. Through power analyses, it is possible to determine the chance that a prespecified effect will be detected. A power of .8 to detect effects (equivalent to a type II error rate of .2) is often seen as acceptable but retrospective analyses of published papers typically show much lower power (Brysbaert & Stevens, 2018). To design sufficiently powered experiments, it is necessary to calculate the power to determine a prespecified effect a priori. The size of the effect of interest also has to be specified and this size needs to be the minimal interesting size that would support the hypotheses (Dienes, 2014). The power can then be approximated using analytical formula employed in free software such as G*Power (Faul et al., 2007) or by using Monte Carlo simulations (e.g. Caldwell & Lakens, 2019; Green & MacLeod, 2016).

Next to the number of participants and the experimental design, the power to detect effects depends also on the statistical methods. To achieve sufficient power, it is thus important to employ best possible procedures to maximize said power to detect effects. While there is no perfect method, there have been many recent advancements in and discussions about statistical analyses. Traditionally, multifactorial designs were commonly analyzed with ANOVAs. More recently, linear mixed models and general mixed models became more widely spread, in part due to the increased availability of computing power and easy to use packages (Bates, Mächler, et al., 2015). The general advantages of mixed models are well documented but there exists some debate on the choice of specific parameters, especially the choice of random slopes (Barr et al., 2013; Bates, Kliegl, et al., 2015; Brauer & Curtin, 2018; Matuschek et al., 2017). For overviews of the theoretical background and applications of mixed models, see for example Baayen et al. (2008), Barr et al. (2013), Brauer and Curtin (2018), or a more beginner-friendly practical tutorial of Winter (2013).

In short, mixed models employ both random and fixed effects. There is no clear definition of random and fixed effects. Typically, random effects are expected to generalize to a more general sample (e.g., generalizing data of randomly chosen participants to the general population). Fixed effects include all items of interest (e.g., all treatments of interest in one study). Each random effect can furthermore be associated with random intercepts and random slopes. While random intercepts account for baseline differences, random slopes account for possible differences in the interaction of fixed and random effects. Some of the main advantages of mixed models are the treatment of multiple crossed or nested random effects, the possible inclusion of multiple covariates, and the possibility to analyze unbalanced and partially missing data, all while achieving higher statistical power. Moreover, mixed models offer possibilities to analyze the progress during the time course of an experiment (Baayen et al., 2008; Mirman et al., 2008; Winter & Wieling, 2016). The change over time can be of primary interest but also an important covariate.

While mixed models are superior to traditional statistical tests, there are drawbacks to null hypothesis significance testing and another (not incompatible) approach, Bayesian hypothesis testing, is recently gaining popularity (Andrews & Baguley, 2013; van de Schoot et al., 2017; Wagenmakers, 2007). Due to advantages and disadvantages of each approach, there is no clear consensus on the best methods to use. Moreover, both approaches can also be seen as complementary instead of competing (Dienes, 2014; Wagenmakers, 2007). A drawback to null hypothesis significance testing that is not fully solved by Bayesian statistics is the dichotomization of results and the misinterpretation of the non-detection of effects as no effects (Amrhein et al., 2019). Despite the use of state-of-the-art statistical methods, it is thus important to acknowledge the uncertainty of outcomes.

2.6.2 Replicability

The replication rate of previously published results has been identified as rather low in psychology and many other fields (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Next to underpowered studies and a publication bias favoring significant results, a main reason are questionable research practices such as p-hacking (running the experiment or adapting analyses until results are significant, Simmons et al., 2011) and p-harking (hypothesizing after results are known, Kerr, 1998). While wrong conclusions derived from such practices can be identified through repeated replication attempts, best practice should be to eliminate such practices in the first place. Solution attempts include the call to authors to report all data and analyses as well as more scrutiny by reviewers (Simmons et al., 2011). However, as the success of such an approach necessitates transparency by authors, it is important to demonstrate such transparency by open research practices. This includes the preregistration of studies and making data and code publicly available (Nosek et al., 2015).

3 Summary of the State of Research

Mental rotation has been a field of interest in the research of spatial abilities for about 50 years now and still enjoys great popularity. This ability to imagine rotated representations of objects has been shown to resemble the time-path of physical rotations in chronometric tests. Indeed, rotational hand movements have been shown to interact with mental rotation tasks and have been used to enhance mental rotation performance. For psychometric mental rotation tests, a major field of interest are sex differences in performance, the reasons for which are not yet fully understood. The differences between both tests also offer the possibility to further investigate and improve the test design. Regarding these aspects of mental rotation, six studies were conducted. While these topics are interesting by themselves, all are interconnected by the insights gained on mental rotation ability. Thus, instead of ordered by topics, the studies are presented in chronological order, which represents the incremental knowledge gained throughout this thesis.

To date, all mental rotation tests rely on the discrimination between rotated stimuli and distractors. Although they might still be rotated in the mind, the solution of distractor trials does not require mental rotation and they can thus not be analyzed as part of mental rotation performance. However, this means that typically half of the answers and the time spent on obtaining them are not used. By modifying tests in the **second** study, we attempt to make mental rotation more useful in all trials, such that they can all be analyzed for mental rotation ability. This can make research more time efficient, improve statistical power, and remove biases introduced by the distractor trials. Moreover, by investigating differences and similarities between different test designs in the **sixth** study, we aim to further the understanding of the influence of the test design on mental rotation measurements.

Due to the practical implications of the relationship between mental rotation and STEM performance, sex differences in mental rotation performance are an important topic. It has been suspected that they are in part due to the stereotypical nature of the stimulus material. Both manipulating stereotypes as well as explicitly or implicitly inducing emotions have been shown to affect mental rotation performance. Thus, it seems plausible that feelings towards the stimulus material influence the performance, which has been investigated in the **fifth** study. However, sex differences are not observed for all test designs and thus it is not clear whether sex differences in test scores are due to actual differences in mental rotation ability. This has also been further investigated in the **sixth** study.

While mental rotation has been shown to be trainable by manual rotation interventions, the exact mechanisms are unclear. In all employed manual rotation trainings, the rotational movement was linked to a visual representation of a rotating object. Separating the visual rotation from the movement could offer further insight into the similarities between and the processes of manual and mental rotation, which was investigated in the **fourth** study. Whereas mental rotation has this special connection to a rotational movement, there is a more general interaction of cognitive abilities and physical activity, which depends on the types of both activities. However, for aerobic activity, mostly the effect of exercise on cognition has been studied and the other direction has been neglected. By investigating the bidirectional interaction of simultaneous mental rotation and aerobic exercise in the **first** study, we aim to gain more insight into this interaction.

Although these insights into mental rotation are interesting, it is also important to consider the methods used to obtain them. Many cognitive tasks are influenced by short term practice effects and in the second study, we do find improvements even within the testing session. The statistical advancement of linear mixed models allows the combination of traditional group comparisons with a time course analysis. In the **third** study, we explored the usefulness of such a combination for the interpretation of results. Next to statistical advancements, it has been shown that previous scientific results are not as reliable as they should be. The methodological advancements to improve confidence in future results include power analyses, the preregistration of studies, data and code availability, and a cautious discussion of the uncertainty of results independent of significance. Throughout the studies, I have tried to adhere to these practices to the best of my knowledge.

4 First Study: Interactions Between Simultaneous Aerobic Exercise and Mental Rotation[1]

4.1 Goals and Hypotheses

While the effects of aerobic exercise during a cognitive task on the performance of said cognitive task have been extensively studied, it has not been investigated whether cognitive performance during aerobic exercise influences the physical performance. It is the main goal of this study to explore both directions of the interaction of mental rotation and aerobic exercise, namely the possible influences of mental rotation on physical exercise as well as the effect of physical exercise on mental rotation while also analyzing subjective effort by measuring the rating of perceived exertion (RPE).

To this end, we will compare the simultaneous performance of cycling and mental rotation to both isolated cycling at the same power output and isolated mental rotation conditions. According to both the RAH (Dietrich & Audiffren, 2011) and the strength model (Audiffren & André, 2015; Baumeister et al., 1998, 2007), negative effects in both directions are expected, that is, an increased heart rate and slower and less accurate responses for the mental rotation task as well as larger RPEs. Additionally, with increasing exhaustion during the experimental task, we expected these effects to increase over time.

4.2 Method

4.2.1 Participants

For the power analysis, Brysbaert and Stevens (2018) suggest using the typical effect sizes in psychology of $d = 0.4$ or $d = 0.3$. For effect sizes of $d = 0.4$, G*power (Faul et al., 2007) reveals a power of .81 for within-subject comparisons of 41 participants, which should further increase

[1] The results presented in this chapter were published in advance in: Jost, L., Weishäupl, A., & Jansen, P. (2021). Interactions between simultaneous aerobic exercise and mental rotation. Current Psychology. https://doi.org/10.1007/s12144-021-01785-6

with the use of linear mixed models (Barr et al., 2013; Hilbert et al., 2019). With the planned duration of the mental rotation tasks of 30 minutes and an estimated average reaction time of 3s[2], this should also exceed the recommendation of Brysbaert and Stevens (2018) for analyses of reaction times of at least 1600 observations per condition to detect commonly observed very small effect sizes of around *d* = 0.1. For the effect of exercise on cognitive performance, Chang et al. (2012) report effect sizes of d=0.10 during exercise, moderated by intensity (moderate intensity with *d* = 0.19), task type (executive function as the most comparable category with *d* = 0.26), and fitness level (moderate fitness level with *d* = 0.01). As we expect much lower variance in physiological measures within participants, the estimated effect sizes of *d* = 0.4 and *d* = 0.1 seem appropriate.

Accordingly, a total of 41 German sport students (22 women, 19 men) participated in this study. Participants were recruited through a seminar as part of which the maximal performance tests were performed and received additional study credit. Participants chose the day of the week and the time of the day on a first come, first served basis. Starting times were available from 9:00 to 12:00 and 13:45 to 17:30 in 45-minute slots from Monday to Friday. Two additional students wanted to participate but had to withdraw prior to the first session, one due to illness and one due to timing conflicts. All participants reported no limitations regarding physical exercise or digitally presented cognitive tasks. Participants were instructed to continue with their usual eating, sleeping, and training habits but not to perform additional training on the same day prior to testing. Self-reported demographic data is summarized in table 1.

[2] As only correct answers on rotated trials are analyzed, we expected such a trial once in about 10s or about 180 such trials per participant over 30 minutes.

Table 1

Participants' Data.

Variable	Men	Women
n	19	22
Age (years)	22.7(2.35)	20.1(1.58)
Height (cm)*	181.1(6.35)	166.2(5.16)
Weight (kg)*	79.3(13.19)	60.0(7.10)
Physical activity (h/week)*	9.8(4.12)	7.0(4.34)
Handedness	5.4(6.79)	8.4(2.65)
Maximal power (W)*	273.4(39.79)	178.9(23.35)
Relative maximal power (W/kg)*	3.5(0.56)	3.0(0.45)
Maximal heart rate (bpm)	188.1(10.16)	189.8(10.13)
Time until exhaustion (min)*	16.8(3.37)	11.9(2.33)

Note. Mean (SD) of participants anthropometric and physiological data, separated by sex. Significant differences ($p < .05$) by sex are marked by *.

4.2.2 Material

4.2.2.1 Chronometric Mental Rotation Test

Stimulus presentation and response handling were controlled with Presentation® software (Version 20.1 Build 12.04.17, Neurobehavioral Systems, Inc., Berkeley, CA) on a Dell Latitude E7240 Laptop and presented on an external 22-inch Dell 2208WFP monitor, 1366x768, 59Hz, positioned approximately 50cm in front of the handlebars independent of participants' size in the experimental sessions and 50cm in front of a chair in the practice session.

Stimuli were obtained from the stimulus library of Peters and Battista (2008). All 16 models with a rotation around x- and z- axis in 45° steps and mirrored/non-mirrored orientations (a and b orientation in the stimulus library) were used with a checkered pattern on a black background, resulting in a total of 480 different stimuli. On the left side of the screen the model was presented in orientation a, rotated by 30° in x direction and 15° in z direction, such that the base model for x or z rotations was identical. On the right side of the screen, a rotated and mirrored/non-mirrored stimulus was presented. Stimulus pictures were sized 400px times 400px and presented vertically centered and horizontally positioned 300px to the left or right of the center of the screen (see figure 2) until a response was given. Between stimulus pairs in the practice session, participants received feedback for 1000ms (✓- right, ✗- wrong) displayed in the center of the screen at font size 40, in experimental sessions a fixation cross ("+") was shown at the center of the screen for 500ms.

Stimuli were presented in a predefined random order until either 10 minutes (practice session) or 30 minutes (experimental sessions) had passed. Three orders of stimulus pairs were randomly generated using shuffling in Presentation® software, such that all stimuli were shown once in random order before they were shown again. Between two occurrences of the same stimulus pair, at least 30 other stimulus pairs were shown. In the practice session, the same order was used for all participants. However, for the two experimental sessions, the remaining two orders were assigned randomly, counter-balanced between participants and the order of experimental sessions using random permutations in R (R Core Team, 2018).

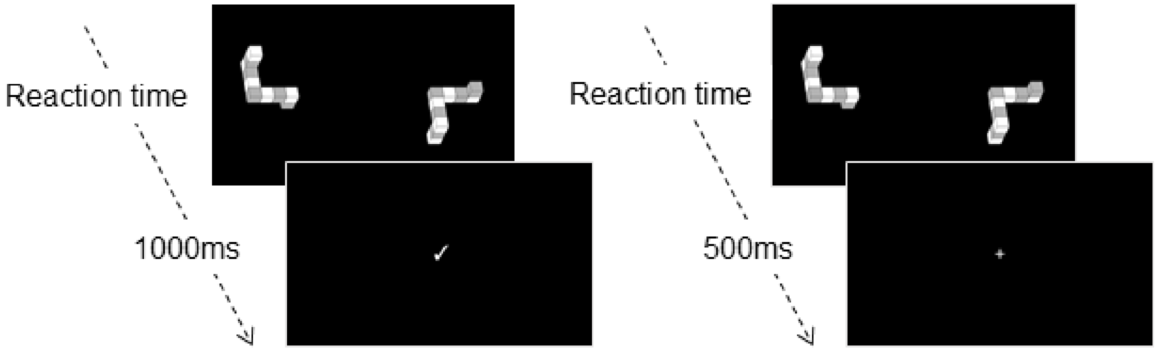
In addition to the stimulus pairs, total trial duration was presented in 15s steps at the top of the screen (horizontally centered, 300px above the center of the screen, font size 48), albeit only updated when the screen changed between stimuli, not while one stimulus pair was shown.

Participants were digitally instructed to press the left mouse button if the stimuli could be rotated into congruence (non-mirrored), and the right mouse button if the two stimuli were mirrored and to answer as quickly and as precisely as possible. Mouse handling was not specified. Thirty-five participants used their right hand and kept it constantly on the mouse in all sessions, whereas six participants used different mouse handling but used similar handling in both experimental sessions involving mental rotation.

Reaction time, accuracy, and stimulus type (model, rotation angle, rotation axis, orientation), as well as time since the start of the trial, were recorded.

Figure 2

Structure of a Mental Rotation Trial.



Note. First, participants performed the mental rotation task with unlimited time until they gave a response. In the practice session, they then received feedback (✓ - right, ✗- wrong) for 1000ms (left figure). In the experimental sessions, they saw a fixation cross (+) for 500ms (right figure). The stimulus pair is an example of a rotation around the x-axis by 180°.

4.2.2.2 Physiological Measures

Cycling was performed on a Cyclus2 ergometer (RBM elektronik-automation GmbH, Leipzig, Germany) allowing control of power independent of pedal cadence with a racing bicycle frame and racing saddle (Cube Peloton, Size S). Saddle height was adjusted to participants individually, such that they were able to cycle comfortably, although handlebar height remained constant for all participants. Participants were instructed to cycle at a cadence of their choice, but at least 50 rpm. If it dropped below 50 rpm, participants were asked to increase their cadence (only applied to one participant in the only exercise condition). Heart rate was measured using a heart rate belt (Polar H7 or Cosmed) and was transferred to a watch (Polar M400) or Cyclus2.

Heart rate was documented every 5 minutes in all experimental sessions and pedal cadence in all experimental sessions involving exercise.

4.2.2.3 Subjective Effort

Both subjective physical effort (physical rating of perceived exertion, pRPE) and subjective cognitive effort (cognitive rating of perceived exertion, cRPE) were measured using the "Anstrengungsskala Sport" (ASS), an RPE scale developed in German by Büsch et al. (2015) ranging from 0-10. The scale was shown to participants and they were asked to name a unique number or word description describing their physical or cognitive effort. In the experimental sessions, RPE measurement followed directly after finishing the exercise/tasks. In the case of simultaneous physical and cognitive tasks, pRPE was measured first.

4.2.3 Procedure

This study employed a cross over design and participants were required to visit the laboratory on four occasions (one pre-examination and practice session and three experimental sessions) a week apart on both the same day of week and same time of day. Twenty-nine participants fulfilled these timing requirements, ten participants differed from this rhythm, but had at least 48h of rest between sessions; two participants only managed to participate on three occasions. Temperature, humidity, light, and sound conditions were not controlled for.

In the pre-examination and practice session, participants were informed about the study design and goals and their individual order of experimental sessions and were also instructed about the study's usage of the ASS. An incrementing stage test was performed to exhaustion with an increase of 30W every 3 minutes. Starting power for women was set to 90W, for men weighing less than 80kg to 120W, and for men weighing at least 80kg to 150W. For one self-reported competitive endurance-athlete, the starting power was increased by a further 30W. This is in line with some commonly applied test protocols with consideration for varying absolute power output by sex, weight, and endurance fitness level, but there exists no consensus on test protocols (Bentley et al., 2007; Faria et al., 2005b). The end of the test was determined by subjective exhaustion (an RPE of 10) despite strong verbal encouragement. RPE was measured in the last 10s of every stage for habituation purposes. Heart rate and pedal cadence were continuously measured and could be seen by participants. Maximal heart rate and maximal power were documented, where maximal power was calculated as the power of the last completed stage plus 5W for every additional 30 seconds completed of the next stage to adjust for the proportion of the last stage (Hopkins et al., 2001). Maximal heart rates exceeded 95% of participants' age predicted peak heart rate (200 - age; Such & Meyer, 2010), indicating that all participants achieved exhaustion objectively.

After finishing the performance test, participants were allowed to rest until they felt ready to continue with the practice session of 10 minutes of mental rotation. To reduce familiarization with the experimental conditions, participants were seated in a chair for the practice session. At the end of the first session, cRPE, age, height, weight, physical activity (the sum of aerobic activity, weight training, and sports games), and handedness according to the Edinburgh handedness inventory (Laterality quotient ranges from -10 to +10; Oldfield, 1971) were measured using a digital questionnaire.

All experimental conditions lasted for 30 minutes. In the exercise only (E) and the mental rotation while exercising (ME) conditions, participants cycled at 60% of their individual maximal power, rounded to 5W. Individual maximal power has shown to be a good predictor of aerobic cycling performance and 60% corresponds to a moderate intensity range, such that all participants could cycle at this intensity for 30 minutes without premature exhaustion. An intensity prescription using blood lactate thresholds was avoided as invasive methods must be used and small variations in intensities should have a limited effect on cognitive performance.

The duration was set to 30 minutes, as this is generally used to evaluate lactate steady states and should thus be sufficient to reach valid measurements of heart rate and RPE; it also matches or exceeds duration recommendations for aerobic activity (Bentley et al., 2007; Faria et al., 2005b; Hohmann et al., 2002). In the mental rotation only (M) condition, participants sat on the bicycle frame without cycling. Participants had no knowledge of their power output, pedal cadence, or heart rate, but were informed that power was kept constant over the duration thereof. In the E condition, participants were informed every 5 minutes about the time, otherwise there was no interaction between conductor and participant and conductors stayed out of participants' visual field.

Additionally, before the last experimental session, baseline heart rate was measured after adjusting the ergometer and participants stayed seated but rested on the ergometer for 30 seconds.

4.2.4 Study Design

Three experimental conditions were used: exercise only (E), mental rotation only (M), and mental rotation while exercising (ME). The M condition was always performed in the second experimental session, whereas E and ME were performed in random order in the first and third experimental sessions, counter-balanced between participants and the random orders used for mental rotation, such that all participants had two weeks of rest between physical tests and one week of rest between cognitive tests.

For cognitive performance, dependent variables were reaction time, accuracy, and cRPE. For physical performance, dependent variables were pedal cadence, heart rate, and pRPE. Independent variables were condition (M and ME for cognitive performance and E and ME for physical performance), time (since start of the session), the interaction between condition and time, sex, and for cognitive performance also the angle of rotation.

4.2.5 Statistical Analysis

Statistical analysis was performed with linear mixed models using the lme4 package (Bates, Mächler, et al., 2015) in R (R Core Team, 2018). Model parameters were estimated by maximum likelihood estimation. P-values were obtained by using likelihood ratio tests to test for improvement of model fit by the fixed effect of interest and compared to a significance level of 0.05. Confidence intervals were calculated using parametric bootstrapping with 1000 simulations. A visual inspection of residual plots did not reveal any deviations from homoscedasticity or normality in any model.

Where possible, we report both the unstandardized effect sizes and confidence intervals calculated by using parametric bootstrapping with 1000 simulations in line with the recommendations of Baguley (2009) and Pek and Flora (2018). While standardized effect sizes are routinely used for power analysis and meta analyses, unfortunately there is not a consensus regarding how to compute standardized effect sizes in linear mixed models (Feingold, 2009; Hedges, 2007; Rights & Sterba, 2019). Nevertheless, linear mixed models offer several advantages over traditional ANOVAs. For example, linear mixed models allow the simultaneous analysis of by-participant and by-item variances, thus eliminating the need to average over participants or items, while also facilitating the analysis of unbalanced data and achieving higher statistical power (Barr et al., 2013; Hilbert et al., 2019).

Model building was based on the research of Barr et al. (2013) and Bates, Kliegl, et al. (2015), starting with a model with random intercepts and slopes for every appropriate fixed effect and subsequently reducing the model complexity by dropping non-significant variance components. Non-significant fixed effects were further removed from the model, such that non-significant effects were tested for an improvement of model fit by inclusion in the resulting model, while significant effects were tested for worsening of model fit by exclusion of the effect. The resulting models for each parameter are described in the results section.

As there is ongoing discussion about model selection based on theory or data or preferring complex instead of simpler models, we expect future research to cast doubt on the optimality of the currently suggested models. Although models without random slopes seem too anti-conservative (Barr et al., 2013), we report the results of these simplest models for easier comparison.

As it is not clear whether RPE values should be treated as interval or ordinal (e.g. Bishop & Herron, 2015), all statistical analyses of RPE values have been conducted twice, with one treating the values as ordinal and one as interval. No differences in significance were found in any analysis and we report only the results obtained by treating the values as interval scaled.

4.3 Results

4.3.1 Physical Performance

4.3.1.1 Pre-Examination Data

The results of the performance test and questionnaire are presented in table 1. Most notably, we found differences by sex in physical activity and time until exhaustion in the performance test with longer test duration for men.

4.3.1.2 Heart Rate During Exercise

For the analysis of heart rates, the model building resulted in a model with a random intercept and random slopes for condition and time by participant. Condition, sex, time, and the interaction between condition and time were analyzed as fixed effects; in doing so, significant effects were found for sex, time, and condition*time but not for condition (see table 2 for inferential statistics and figure 3 for the descriptive comparison of conditions over time). Heart rate increased significantly with time and women showed significantly higher heart rates; however, differences between conditions were not significant. The point estimate of less than 1bpm indicates no meaningful overall difference between conditions. The analysis of the interaction showed a significantly lower intercept and larger increase in heart rate in the E condition, while differences at our measured time points were only significant at the 30-minute mark.

Table 2

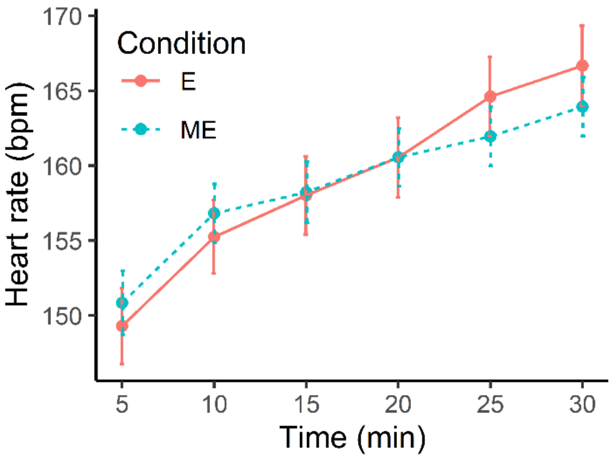
Statistical Analysis of Heart Rate During Exercise.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	141.35	2.89			135.62, 147.24
Condition (ME-E)	-0.03	1.19	$\chi^2(1)=0.00$	.98	-2.44, 2.42
Time (30 min)	17.32	1.14	$\chi^2(1)=78.0$	<.001	15.04, 19.63
Sex (female-male)	11.49	3.49	$\chi^2(1)=8.9$	.003	4.23, 18.67
Condition*time			$\chi^2(1)=33.2$	<.001	
Intercept (ME-E)	3.14	1.30			0.47, 5.71
E*time	20.08	1.22			17.63, 22.42
(ME-E)*time	-5.43	0.92			-7.17, -3.63

Note. Intercepts in this model represents the estimate for E condition at time 0 for male participants. SE- standard error, CI – confidence interval, E - only exercise condition, ME – combined mental rotation and exercise condition.

Figure 3

Heart Rate During Exercise.



Note. Line plots showing mean heart rate as a function of time for both exercise conditions. Error bars show standard error. E - only exercise condition, ME – combined mental rotation and exercise condition.

Table 3

Statistical Analysis of Heart Rate Without Exercise.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	84.36	2.04			80.01, 88.28
Condition (baseline-M)	0.13	1.89	$\chi^2(1)=0.00$	.95	-3.76, 3.76
Time (30 min)	2.48	1.34	$\chi^2(1)=3.31$	.07	-0.09, 4.99
Sex (female-male)	5.17	4.02	$\chi^2(1)=1.60$	.21	-2.31, 13.50

Note. Intercepts in this model represents the estimate of the grand mean. Values for intercept and sex are from the model using both M condition and baseline. SE- standard error, CI – confidence interval, M- only mental rotation condition.

4.3.1.3 Heart Rate Without Exercise

In the analysis of time in the M condition, random intercepts and slopes for time by participants were used. To compare with baseline and analysis of sex, random intercepts and slopes for condition by participants were used. Condition, time, and sex were analyzed as fixed effects and no significant effects were found (see table 3). However, in a model without random slopes by time the increase of heart rate by time was significant (p = 0.01).

4.3.1.4 Pedal Cadence

For the analysis of pedal cadence, the model building resulted in a model with random intercepts and random slopes for condition and time by participants. Condition, sex, time, and condition*time were analyzed as fixed effects and significant effects were found for condition, time, and condition*time (see table 4 for inferential statistics and figure 4 for the descriptive comparison of conditions over time). Pedal cadence was significantly lower in the ME condition by an estimated 8.32 rpm and increased significantly over time. The analysis of the interaction between time and condition revealed a significant increase with time only in the E condition, but not in its ME counterpart. Pedal cadence did not differ significantly at 5 minutes, but all later times showed significantly higher cadences in the E condition.

Table 4

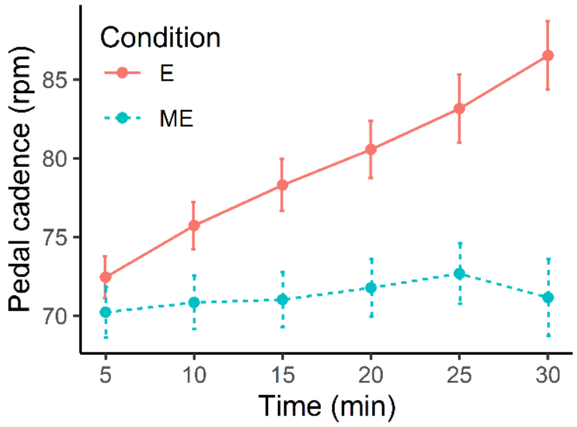
Statistical Analysis of Pedal Cadence.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	70.06	1.68			66.78, 73.48
Condition (ME-E)	-8.32	1.85	$\chi^2(1)=16.21$	<.001	-11.82, -4.64
Time (30 min)	9.20	1.59	$\chi^2(1)=24.58$	<.001	5.99, 12.19
Sex (female-male)	-1.32	2.62	$\chi^2(1)=0.23$	.63	-6.69, 3.84
Condition*time			$\chi^2(1)=77.59$	<.001	
Intercept (ME-E)	0.25	2.06			-4.01, 4.51
E*time	16.60	1.81			12.70, 20.22
(ME-E)*time	-14.76	1.58			-17.74, -11.49

Note. Intercepts in this model represents the estimate for E condition at time 0. SE- standard error, CI – confidence interval, E - only exercise condition, ME – combined mental rotation and exercise condition.

Figure 4

Pedal Cadence During Exercise.



Note. Line plots showing mean pedal cadence as a function of time for both exercise conditions. Error bars show standard error. E - only exercise condition, ME – combined mental rotation and exercise condition.

4.3.1.5 Physical Subjective Effort

For the analysis of pRPE, the model building resulted in a model with only a random intercept by participant. Condition and sex were analyzed as fixed effects and neither showed significant differences (see table 5 for inferential statistics and figure 5 for the descriptive comparison). The estimated difference of pRPE between conditions of -0.05 suggests no meaningful difference. For the comparison of sexes, the estimated difference of 0.15 might need further investigation.

4.3.2 Mental Rotation

Outliers were determined by a deviance of more than three standard deviations from the mean reaction time of all stimulus pairs with the same rotation angle and were excluded from all analyses. Because angular disparity is not defined for mirrored responses in cube figures (R. N. Shepard & Metzler, 1971), only non-mirrored stimulus pairs were analyzed and reaction time was also only analyzed on correct responses.

4.3.2.1 Reaction Time

Model construction resulted in a model with random intercepts and slopes for condition, time and degree by participant, and random intercepts and slopes for time and degree by model. Time, condition, time*condition, angular disparity, and sex were analyzed as fixed effects and significant differences were found for angular disparity and time (see table 6 for inferential statistics and figure 6 for the descriptive comparison of conditions and angular disparity). Reaction time improved significantly over time and significantly increased by degree. The estimates between both condition (27.45ms faster in the M condition) and sexes (females 0.52ms slower) are less than 1% of the average reaction time and suggest no meaningful differences.

Table 5

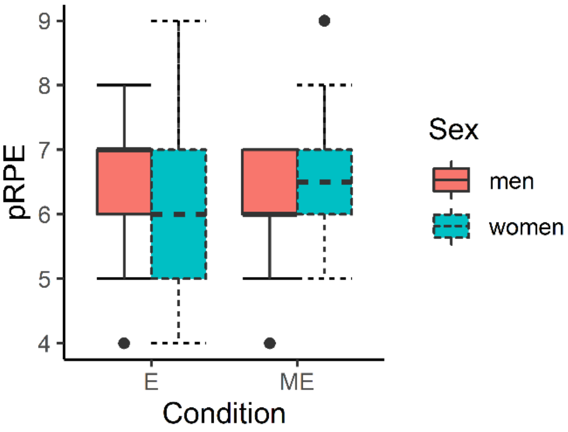
Statistical Analysis of pRPE.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	6.35	0.15			6.06, 6.64
Condition (ME-E)	-0.05	0.16	$\chi^2(1)=0.10$	.75	-0.36, 0.24
Sex (female-male)	0.15	0.30	$\chi^2(1)=0.25$	.62	-0.47, 0.72

Note. Intercepts in this model represents the grand mean. SE- standard error, CI – confidence interval, E - only exercise condition, ME – combined mental rotation and exercise condition.

Figure 5

PRPE After Exercise.



Note. Box plots showing pRPE after exercise conditions for both sexes. Whiskers are restricted in length to 1.5*IQR. E - only exercise condition, ME – combined mental rotation and exercise condition.

Table 6

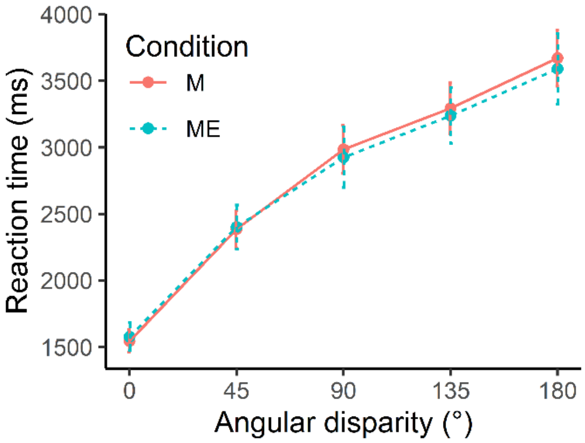
Statistical Analysis of Reaction Time.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	2310.59	127.06			2044.21, 2559.32
Time (30 min)	-785.99	92.14	$\chi^2(1)=37.61$	<.001	-961.93, -601.83
Degree (100°)	1057.27	85.80	$\chi^2(1)=55.91$	<.001	895.61, 1227.55
Sex (female-male)	0.52	162.19	$\chi^2(1)=0.00$	.99	-315.69, 338.95
Condition (M-ME)	-27.45	128.72	$\chi^2(1)=0.05$	.83	-266.64, 248.64
Condition*time			$\chi^2(1)=0.65$	.42	

Note. Intercepts in this model represents the estimate at time 0 and unrotated stimuli. SE- standard error, CI – confidence interval, M - only mental rotation condition, ME – combined mental rotation and exercise condition.

Figure 6

Reaction Time of Mental Rotation Tasks.



Note. Line plots showing mean reaction time of mental rotation trials as a function of angular disparity for both cognitive conditions. Mean reaction time is calculated for all non-mirrored and correctly answered trials of every participant and then averaged over all participants. Error bars show standard error. M - only mental rotation condition, ME – combined mental rotation and exercise condition.

4.3.2.2 Accuracy

Accuracy was analyzed by a general linear mixed model, which used a binomial distribution with random intercepts and slopes for time and degree by participant and random intercepts and slopes for time by model. Time, condition, time*condition, angular disparity, and sex were analyzed as fixed effects and significant differences were found for angular disparity and time (see table 7 for inferential statistics and figure 7 for the descriptive comparison of conditions and angular disparity). Accuracy improved significantly over time and decreased significantly by degree. The point estimates for differences between condition and sex correspond to average changes in accuracy of 0.1% and 1.6% of trials, respectively. For the comparison of conditions in particular, this suggests no difference exists.

4.3.2.3 Cognitive Subjective Effort

For the analysis of the cRPE, the model building resulted in a model with only a random intercept by participant. Condition, sex, and their interaction were analyzed as fixed effects and we found significant differences for both condition and sex, but the interaction was not significant (see table 8 for inferential statistics and figure 8 for the descriptive comparison). Subjective effort was significantly lower in women compared to men and in the M condition compared to ME, both by approximately one value on the scale.

Table 7

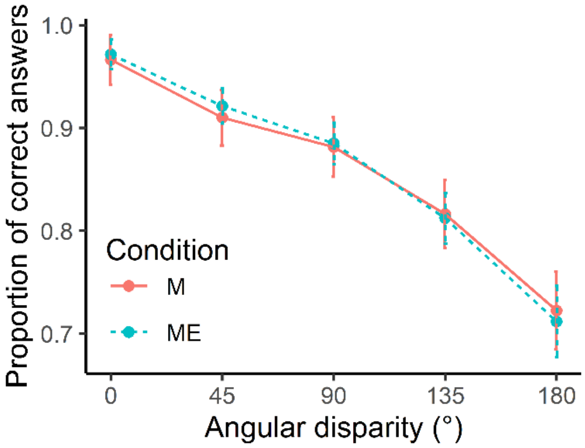
Statistical Analysis of Accuracy.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	3.56	0.23			3.14, 4.01
Degree (100°)	-1.48	0.11	$\chi^2(1)=68.15$	<.001	-1.71, -1.27
Time (30 min)	0.52	0.16	$\chi^2(1)=9.89$	.002	0.20, 0.82
Condition (M-ME)	-0.03	0.04	$\chi^2(1)=0.53$	.47	-0.12, 0.06
Sex (female-male)	-0.48	0.32	$\chi^2(1)=2.22$	.14	-1.04, 0.10
Condition*time			$\chi^2(1)=0.02$	.89	

Note. Intercepts in this model represents the estimate at time 0 and unrotated stimuli. SE- standard error, CI – confidence interval, M - only mental rotation condition, ME – combined mental rotation and exercise condition.

Figure 7

Accuracy of Mental Rotation Tasks.



Note. Line plots showing mean accuracy of mental rotation trials as a function of angular disparity for both cognitive conditions. Mean accuracy is calculated for all non-mirrored trials of every participant and then averaged over all participants. Error bars show standard error. M - only mental rotation condition, ME – combined mental rotation and exercise condition.

Table 8

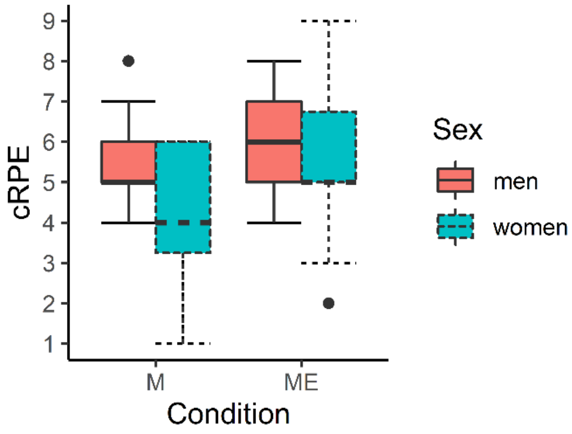
Statistical Analysis of cRPE.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	6.29	0.30			5.71, 6.87
Condition (M-ME)	-0.96	0.25	$\chi^2(1)=12.26$	<.001	-1.49, -0.45
Sex (female-male)	-0.92	0.37	$\chi^2(1)=5.88$	.015	-1.63, -0.22
Condition*sex			$\chi^2(1)=0.12$	.73	

Note. Intercepts in this model represents the estimate in the base model for ME condition and male participants. SE- standard error, CI – confidence interval, M - only mental rotation condition, ME – combined mental rotation and exercise condition.

Figure 8

CRPE After Mental Rotation.



Note. Box plots showing cRPE after cognitive conditions for both sexes. Whiskers are restricted in length to 1.5*IQR. M - only mental rotation condition, ME – combined mental rotation and exercise condition.

4.4 Discussion

The aim of the study was to explore the possible influences of simultaneous cognitive tasks on aerobic exercise and to further analyze the effects of exercise on complex cognitive function in the form of a mental rotation task. Our main results reveal no significant effects of neither simultaneous cognitive tasks on objective (heart rate) or subjective (RPE) physical effort nor of simultaneous exercise on reaction time or accuracy in cognitive performance. However, we identified lower cadence during the simultaneous mental rotation tasks; this was also stable in time compared to an increase in cadence during exercise control. Furthermore, our results demonstrated increased cognitive effort during exercise.

Since the increase in pedal cadence is linked to an increased physiological demand even at constant power, a lower heart rate should have been found during the cognitive tasks. Because the heart rate was comparable in the isolated exercise and the simultaneous exercise and mental rotation condition, this indicates that additional physiological effort was necessary in the combined condition due to the cognitive task. Similarly, the increased cognitive effort during exercise indicates that additional cognitive effort was necessary in the combined condition. Both can be interpreted in support of the RAH model (Dietrich & Audiffren, 2011) and its hypothesized reverse direction, that is, brain activity is not only caused by exercise, but contrary brain activity, as associated with cognitive function, also impacts exercise; as additional effort was necessary to maintain this performance, the tasks negatively impacted each other.

4.4.1 Effects on Physical Performance

The result of the lower cadence during the cognitive task could be explained by the assumption that the mental rotation task might stabilize cadence by serving as a distraction, while our control condition might subjectively be too boring and subsequently lead to increased cadence. While there is a linear relationship between power output and effort measured as heart rate or oxygen uptake (Arts & Kuipers, 1994), there is evidence that this relationship is modulated by pedal cadence (Coast & Welch, 1985; Faria et al., 2005a; MacIntosh et al., 2000). The optimal cadence, that is, the cadence which produces the lowest effort at a given power output, increases linearly with power output (Coast & Welch, 1985; MacIntosh et al., 2000) but might also be higher for trained cyclists (Faria et al., 2005a). Both higher and lower than optimal cadences would lead to higher effort being exerted at given power outputs, that is, a higher physiological demand despite the same energetical demand. However, without instruction or sufficient experience, people may be unable to ascertain their optimal pedal cadence and instead choose a higher than optimal cadence.

Research from Coast and Welch (1985) and MacIntosh et al. (2000) suggests that cadences of 70 rpm, as adopted on average in our ME condition, become optimal at around 250W. As 250W is larger than any power implemented in our experimental conditions, this suggests higher than optimal cadence in both E and ME conditions. Because cadence was even higher in the E condition, one would expect even higher effort, both objectively and subjectively. As efforts were comparable between conditions, our results thus suggest that performing cognitive tasks while cycling has similar effects on heart rate and subjective effort as increasing cadence by a further 5-10 rpm above the optimal cadence, but the relationship between heart rate and subjective effort remains unchanged. Further research with better control of physiological demand is necessary to validate this relationship.

While the increase of heart rate over time can be explained by the differing cadence and the cardiovascular drift (the increase of heart rate over time during prolonged exercise, see e.g. Coyle & González-Alonso, 2001), we have found a possible increase in heart rate caused by mental rotation alone in the resting condition. Thus, it is possible that mental rotation during aerobic exercise also causes a heart rate increase over time, just not by as much as in our exercise control condition due to the increasing cadence. However, due to the cardiovascular drift, the influence of the increase in cadence and the cognitive task on heart rate over time cannot be isolated.

Furthermore, we have compared resting cognitive heart rate to our baseline, which was measured under testing conditions in anticipation of exercise and found no difference. This suggests that even very light movement, like being strapped to a bicycle frame, the anticipation of exercise, or general nervousness from being tested, has a similar effect on heart rate as cognitive activity.

Our results suggest, in support of the reverse direction in the RAH model (Dietrich & Audiffren, 2011), that a combination of cognitive with aerobic activity increases physical effort and thus might also affect physiological adaption, but an isolated effect of cognitive activity on physiology is unlikely. As the relationship between subjective and objective effort does not change, it seems likely that cardiovascular adaptation remains steady if similar effort is invested. As workload might be lower in a combined activity of similar effort, skeletal muscular adaptations, while not being the primary target of aerobic training, might be influenced.

4.4.2 Effects on Cognitive Performance

We have found no effect of acute physical exercise on cognitive performance and no difference in improvement over the course of the units; however, we have discovered increased subjective cognitive effort. The measurement of cRPE might be influenced by the preceding measurement of pRPE in the ME condition, although we have found no correlation between physical and cognitive effort.

While reaction time and accuracy alone show no difference between conditions, nor any difference in temporal behavior, in contrast to the results of Lambourne and Tomporowski (2010) and Chang et al. (2012), the additional measurement of subjective effort allows for further interpretation. Subjective cognitive effort suggests that the mental rotation task was demanding but not maximal; after investing more effort, performance could be upheld under the more strenuous ME condition. While the assumption of a full depletion of self-control resources in the strength model (Audiffren & André, 2015; Baumeister et al., 1998, 2007) might not have been fulfilled, results can be interpreted in support of the RAH model (Dietrich & Audiffren, 2011). However, further objective measurements of effort and cognitive tasks that deplete neural resources should be employed in future experiments to support this claim.

4.4.3 Sex Differences

We have found increased heart rate in women during exercise and a lower cRPE in mental rotation, but no significant increase in pRPE. Increased heart rate can be explained by significant sex differences in performance test duration with lower durations and larger increases relative to maximal power leading to higher achieved maximal power. The missing concurrent increase in pRPE needs to be investigated further but could be explained by the choice of an overall lower RPE of the female participants, which can also be seen in the lower cRPE in cognitive tasks.

While the non-differences in mental rotation performance are in line with the results of Jansen-Osmann and Heil (2007b), the lower cRPE is in contrast to the results of Campbell et al. (2018), who found increased cognitive effort in women using pupillometry. However, the different results could be explained by the varying measurement methods for the analysis of cognitive effort (subjective rating in this study in contrast to the measurement of cognitive strain in the study of Campbell et al., 2018). Further research is necessary to analyze the possible sex differences in selecting RPE and the potentially lower subjective effort for women in mental rotation tasks.

4.4.4 Practical Implications

The results indicate that increased effort, both physiologically and cognitively, is required during combined physical and cognitive work to maintain performance. Practical applications of a combination of aerobic exercise and cognitive tasks are for example learning during endurance training for athletes or treadmill and cycle desks to promote light physical activity during office work. While the performance could be maintained in our experiment, the increased effort could hinder adaptation of such a combination where long-term motivation is an issue. Moreover, further research is necessary if and to what extent the identified effects apply to highly fit individuals and low intensity exercise.

4.5 Limitations

It is possible that our results are restricted to specific parameters of both physical and cognitive performance and to the tested population of sport students. Limitations of the study also include possible changes in cognitive performance due to differences in upper body stability and in room conditions caused by physical exercise, as well as constant noise from the ergometer, which could be distracting but also drown out ambient noise. Although participants were asked to maintain their training, eating, and sleeping habits and mostly sustained their weekly and daily rhythms, these parameters were not explicitly monitored. Results could also be influenced by differences in emotional wellbeing (level of stress, anxiety, or depression), general fitness level, and time of testing between participants. The incentive to participate in this study was also through a seminar, which might affect results differently than the commonly used incentives of only study credit, monetary incentives, or wholly voluntary participation.

A technological limitation is the control of pedal cadence. As our participants did choose different cadences in the two conditions, the conclusions drawn rely on the relationship between cadence and physiological effort. While it is impossible to keep both power and pedal cadence constant at the same time and some control of power seems necessary, more comparable pedal cadences would be preferable to isolate the effects of cognitive tasks on physical effort.

Methodologically, it is possible that the transfer from the practice session to the rested condition was higher, as both were performed without aerobic exercise, although different seating conditions were used. Additionally, as RPE was only measured after the unit, time effects could only be observed in the objective parameters.

4.6 Conclusion

In conclusion, we evidence that during simultaneous aerobic exercise and mental rotation in support of the RAH model (Dietrich & Audiffren, 2011) subjective cognitive effort increases, while cognitive performance can be maintained. Moreover, in support of a possible reverse direction of the RAH model, physiological effort, mediated by pedal cadence, also increases. While the RAH model offers a rather simple explanation for the detected effects, further research is necessary to align the model with contradicting evidence. Measurements of effort might offer some insight into performance (non-)differences. Furthermore, for all future work employing simultaneous exercise and cognitive tasks we suggest employing a physical control condition in addition to the cognitive control condition.

However, to reinforce models like the RAH, brain image studies while participants perform a cognitive as well as a physical task must be conducted. Currently, fNIRS seems to be the best neuroimaging method to obtain such measurements, but more advancements in this technology are necessary to accurately measure and interpret cerebral oxygenation and the hemodynamics of combined aerobic and cognitive tasks. Next to technological limitations and methodological concerns, possible complications also include physiological changes (e.g., heart rate), movement artifacts, and influences of posture (Herold et al., 2018).

5 Second Study: Analyzing All Trials in Chronometric Mental Rotation Tests[3]

In the first study, we employed a classical chronometric mental rotation test without further investigating the design itself as part of the main results. However, further reflecting on the test itself has led us to thinking about possible improvements and systematizations about the test. For now, the about half of all trials (the mirrored trials) are excluded from analyses. Furthermore, there does not seem to be a systematic usage of different rotational axes except for investigations of differences between rotations in the picture plane and in depth.

5.1 Goal and Hypotheses

To overcome the shortcoming of the classical chronometrical mental rotation tests we present a novel approach, in which participants are shown three figures instead of two. A similar approach has already been used by Wohlschläger and Wohlschläger (1998) and by two subsequent studies from Wohlschläger (2001) and Krüger and Krist (2009) while investigating the common processing of manual and mental rotation. However, because they were not interested in the development of the design, they did not differ in their analysis between answer alternatives. They found that overall reaction time demonstrated an increase with increasing angular disparity. In the design presented here we will call two figures the base figures, which are mirrored to each other, and the third the stimulus figure, which is a rotated version of one of the base figures. Instead of a same/different decision, participants are tasked to find out if the stimulus can be rotated into congruence with one or the other base figure. As the base figures are mirrored to each other and not mirror-symmetrical, the stimulus is congruent to one and only one base figure. Thus, the rotation into congruence is independent of stimulus orientation.

[3] The results presented in this chapter were published in advance in: Jost, L., & Jansen, P. (2020). A novel approach to analyzing all trials in chronometric mental rotation and description of a flexible extended library of stimuli. Spatial Cognition & Computation, 20(3), 234–256. https://doi.org/10.1080/13875868.2020.1754833

The main goal of this study is to evaluate the new proposed mental rotation paradigm. While in theory the task at hand can be solved by ignoring one base figure and proceeding as in the case of two figures, we believe that presentation of a unique 0°-condition for all stimuli will show the same behavior for both same and different stimuli (or matching the left or the right base figure in this study). Moreover, we believe that for all stimuli the same mental transformations as in the two-figure case are performed and we expect to find the well-known mental rotation effects: An increase of reaction time and a decrease of accuracy with increased angles of rotation (R. N. Shepard & Metzler, 1971). Furthermore, we predict that participants improve over time with larger improvements for larger degrees. By design, we expect these effects not to differ by the side of the correct answer, although we presume a possible effect due to the left/right arrangement in accordance with the reading direction of our participants. Because of the geometric considerations of the rotation axes presented in the method section, we hypothesize to find no differences of the aforementioned effects between rotation axes. As a secondary hypothesis, we expected to find better performances for participants with previous experience in chronometric mental rotation experiments. At least, short-term benefits for mental rotation training on mental rotation tasks have been shown so far (Meneghetti et al., 2017). Due to the small sample size and no or only small sex differences in chronometric mental rotation tests for cube figures (e.g. Jansen-Osmann & Heil, 2007), sex was not regarded as an additional factor. In addition to the empirical evaluation of the test, we have created an extended, flexible library of stimulus material and describe its properties and utilization.

5.2 Method

5.2.1 Participants

A total of 41 German students (32 females, 9 males) participated in this study due to study credit. On average, participants were 21.3 years old (SD=1.8). In our sample, we included participants with self-reported previous experience in mental rotation experiments employing the two-figure design (24 with experience, 17 without). For appropriate power, Brysbaert and Stevens

(2018) recommend at least 1600 observations per condition independent of effect sizes of other work. As we expected to exceed an average of 40 stimuli per participant per condition in our design, this should yield high power for the main hypotheses. Supporting this we found 14447 usable measurements in the smallest dataset equivalent to 1805 observations per condition for the triple interactions of interest. For the secondary hypotheses, Meneghetti et al. (2017) report large effect sizes of d>1 for transfer within chronometric mental rotation. G*power (Faul et al., 2007) shows a power of .80 for a between-subjects t-test at an effect size of d=0.8.

5.2.2 Rationale for Figure Layout

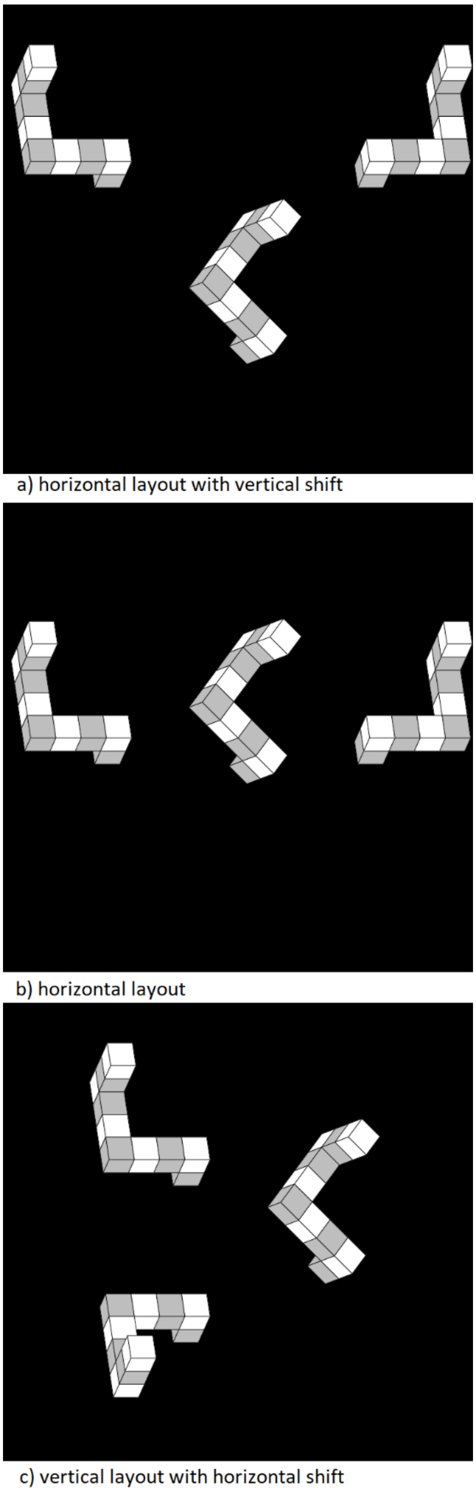
Different layouts are possible to achieve the proposed mental rotation task with three figures with some possibilities presented in figure 9 and many more possible layouts not depicted. We have decided to present the three figures arranged as in figure 9a and 12 for three following main reasons: First, the usage of screen space allows larger images while avoiding overlapping. Second, the layout provides congruence between hand position, response key layout, and figure placement (left/right). Third, the optical distances between stimuli and base figures are small and comparable, while also keeping discrimination between base and stimulus figures clear for participants.

5.2.3 Geometric Considerations of Axes and Mirroring Planes in Three Dimensions

The necessity for the development of the new test presented here is due to the incongruence of mirrored stimuli. In this section, we will review this incongruence and describe geometrical properties of mirroring planes and rotation axes, which we have used as the basis for our design. For the most part, we only consider the natural planes and axes for mirroring and rotation, that is, the xy-, xz-, and yz-plane and the x-, y-, and z-axis. For the naming of the axes, we have used the convention that the vertical axis is the z-axis with larger values for higher points, the horizontal axis is the x-axis with larger values for points further to the right and the depth axis is the y-axis with larger values for points further back. If these axes pass through the center of an object with the front pointing towards the observer, the resulting spins are known as pitch for the rotation around the x-axis, roll for the rotation around the y-axis, and yaw for the rotation around the z-axis. It also must be noted that mirrored stimuli are images of the mirrored figure. As base rotation angles are employed in the 0°-position, these are not actual mirror images. This is demonstrated in figure 10, where images of mirrored figures are shown. While the three-dimensional figures are mirrored to each other, the two-dimensional images of them are not.

Figure 9

Depictions of Different Possible Layouts of Two Base Figures and One Stimulus Figure.



a) horizontal layout with vertical shift

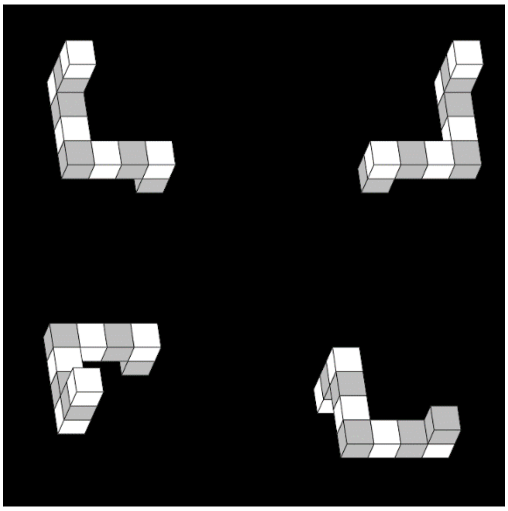
b) horizontal layout

c) vertical layout with horizontal shift

Note. The stimulus figure is always rotated by 45° around the y-axis for demonstration while the base figures are mirrored to each other.

Figure 10

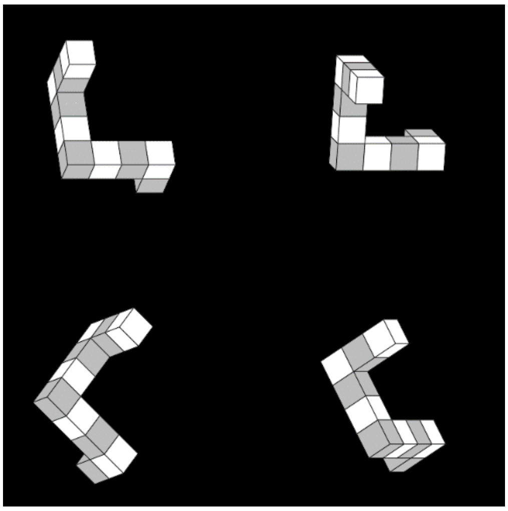
Depictions of Mirror Images Across the Natural Mirroring Planes.



Note. Top left: base model. Top right: Mirroring across the yz-plane. Bottom left: Mirroring across the xy-plane. Bottom right: Mirroring across the xz-plane.

Figure 11

Depictions of Rotation by 45° Around the Natural Rotation Axes.



Note. Top left: Base model. Top right: Rotation around the x-axis. Bottom left: Rotation around the y-axis. Bottom right: Rotation around the z-axis.

5.2.3.1 Incongruence of Mirrored Stimuli

While a mirrored figure is clearly unique in shape, depending on the orientation of the mirroring plane, the mirror image appears at different rotated positions. In figure 10, the images resulting from a mirroring across the xy-plane and the yz-plane differ by a rotation around the y-axis by 180°. More generally, two mirror figures deviate by twice the angle at which the two mirroring planes are rotated to each other. As a result, there exists no unique 0°-position and thus no angular disparity for mirrored stimuli. This problem also occurs in two dimensions, as rotation of mirroring lines also produces rotated stimuli. A case, in which this incongruence might be lessened are for example pictures of human or animal figures, where a clear upright position exists, or mental rotation tasks with egocentric transformations. Rotations around the z-axis, which maintain the upright position, might still cause problems, although one might argue in favor of a unique front facing position, which would also solve this incongruence.

5.2.3.2 Difference and Relationships Between Mirroring Planes and Rotation Axes

As we choose to present two mirrored base figures and by the incongruence no unique way of choosing the mirrored figure exists, the question of choosing a base orientation arises. We propose to use the mirroring plane that optically lies between the two base figures, such that they can more easily be recognized as mirrored shapes of each other. As a result, we use the yz-plane for the mirroring in our layout, while for example in the layout of figure 9c we would use the xy-plane. As we cannot depict a layout in depth on a two-dimensional screen, we ignore mirroring across the xz-plane in this step.

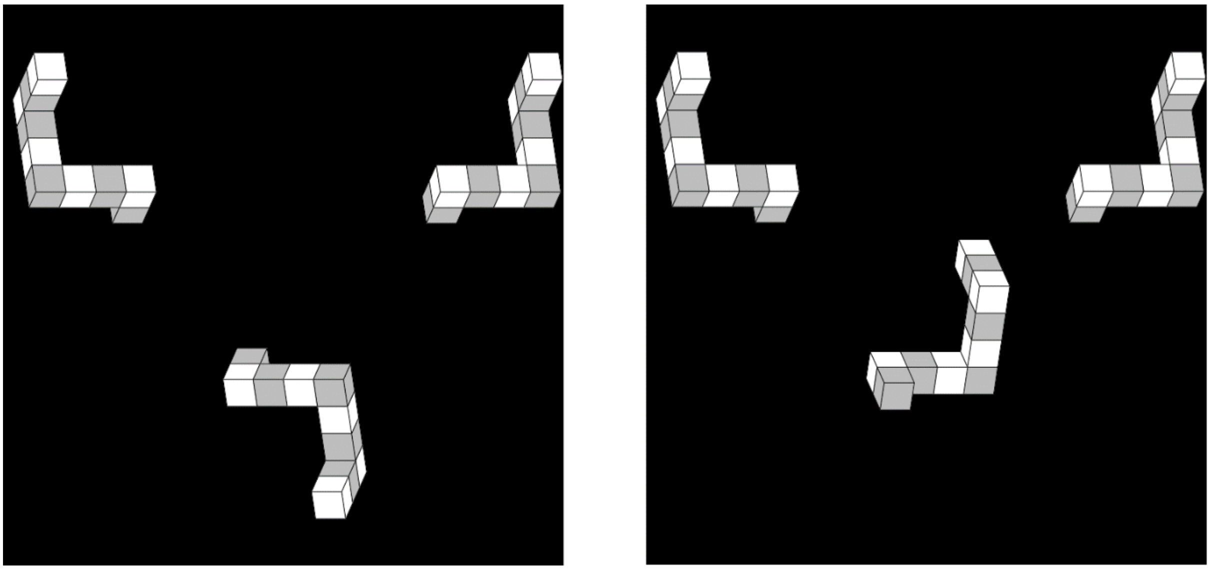
Moreover, mirroring switches the image along one axis while rotation keeps one axis fixed. Mirroring across one plane switches the values along the axis that is perpendicular to that plane, while keeping values constant along the axes that span the plane. Rotation on the other hand keeps the coordinates along the rotation axis constant, while the other two change. As can be seen in figure 10, mirroring across the yz-plane switches the coordinates along the x-axis while keeping the others the same, that is, left and right are reversed. The rotation can be seen in figure 11, where, as the figure is rotated around the x-axis, the left and right parts do not change, but front/back and top/bottom are moved by the rotation. As a result, if a mirroring across the yz-plane and a rotation around the x-axis were used and the rotation axis were known, no rotation strategy would be necessary to solve the task, as a simple left/right comparison would be sufficient. Thus, it seems sensible to choose the rotation axes from the axes that span the mirroring plane. In our case of the yz-plane these are the y- and the z-axis, as we restrict ourselves to the natural axes.

Another problem with a combination of mirroring and rotation is demonstrated in figure 12. A rotation by 180° around the y-axis of the left base figure can easily identified as a mirroring of the right base figure across the xy-plane and similarly a rotation around the z-axis as a mirroring across the xz-plane. Because participants will be informed that only one side can be correct and that the two base figures are mirrored to each other, the task can be solved by exclusion. While this strategy is demonstrated in the case of a rotation by 180°, it is theoretically applicable to all rotation angles, but reaches the natural mirroring planes only at 0° and 180° (and for a rotation of 0° a much simpler solution is available as two figures are identical). As generally larger reaction times are found in the case of mirrored stimuli in classical mental rotation (R. N. Shepard & Metzler, 1971; Voyer & Jansen, 2016), where the same strategy is applicable, this strategy seems too complex to have a major impact on results.

On a further note, the three axes differ in dimension, as rotation around the y-axis is only a rotation in the picture plane, while the x- and z-axis produce rotations in depth, but typically no systematic differences between in-depth rotation and picture-plane rotation are found (R. N. Shepard & Metzler, 1971).

Figure 12

Depictions of a Stimulus Figure That Can be Rotated into the Left Base Figure by a Rotation of 180° Around the y-Axis (Left) and z-Axis(Right).



5.2.4 Material

5.2.4.1 Mental Rotation

Stimulus presentation and response handling were controlled with Presentation® software (Version 20.1 Build 12.04.17, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com) on a Dell Latitude E7240 Laptop (12.5" screen, 1366x768, 60Hz). Participants were seated in front of a desk with the laptop on it and were free to adjust both the laptop and their seating position.

The stimuli were generated from the stimulus library with the parameters given in table 1. Stimulus images were then resized to 400px times 400px and presented according to the layout in figure 1a and 4 until a response was given. The vertical shift of the images was set to 150 pixels (above the center for the base figures and below the center for the stimulus figure) and the horizontal shift to 300 pixels (left and right of the center for the base figures).

Table 9

Parameters for Generation of Cube Figures and Values Used in the Experiment.

Parameter group	Parameter	value
Color options	Background color	black
	Border color	black
	Face color	white, grey
Sizing and formatting	Cube Diameter	70px
	Image size	720px*720px
	File format	png
	Centering	none
Model properties	Base orientations	a,b
	Models	Peters and Battista (2008), 1-16
	Base rotation angles (x,y,z)	-30°,0°,15°
	Angle difference	45°
	Rotation axes	y,z

Between tasks in the practice session, participants received feedback for 1000ms (✓- right, ✗- wrong) shown at the center of the screen at font size 40. In the main session, participants received no feedback and a fixation cross ("+") was shown at the center of the screen for 500ms.

Stimuli were presented until 10 minutes (practice session) or 30 minutes (main session) passed, such that all stimuli were shown once in random order, generated by shuffling in Presentation® software. Afterwards, every stimulus was shown again twice in another random order. Between the first two occurrences of the same stimulus, at least 20 other stimuli were shown. The orientation of the left and right base figure was randomized.

Participants were digitally instructed to press the left mouse button on the touchpad of the laptop with their left index finger if the stimulus could be rotated into congruence with the left base figure. If the stimulus could be rotated into congruence with the right base figure, participants should press the right mouse button with their right index finger. Participants were asked to answer as quickly and as precisely as possible.

Reaction time, accuracy, stimulus type (model, rotation angle, rotation axis, stimulus orientation, base orientation), and time since start of the session were recorded.

5.2.4.2 Demographics

A digital questionnaire was used to collect demographic information. Participants were asked about their previous experience with mental rotation (participants had to indicate if they had or had not participated in other mental rotation experiments before), age (in years), sex/gender (male, female, or other), information about their menstrual cycle, physical and musical activity, and handedness.

5.2.5 Procedure

Participants completed a practice session of 10 minutes and continued with the main session of 30 minutes after a self-paced break. Following the main session, participants answered a digital questionnaire regarding demographic data. Total durations of individual sessions were known to participants before the start of the experiment and were shown again before each session. For the duration of the experiment and the questionnaire, participants were alone in the experimental room.

5.2.6 Statistical Analysis

The accuracy and time of response of each trial were used as dependent variables and the angular disparity, time (since start of the session), the side of the correct answer, the rotation axis, and previous experience of participants were used as independent variables.

Outliers were determined by a deviance of more than three standard deviations from the mean reaction time of all stimulus pairs with the same rotation angle and were excluded from all analyses and reaction time was additionally only analyzed on correct responses. Moreover, as the rotation axis is not well-defined for the 0°-condition, all analyses including axes were calculated only on the rotated stimuli. For all effects not containing axes, the analysis was calculated twice, once on the dataset including the non-rotated stimuli and once on the dataset of only the rotated stimuli, possibly including effects of axes as moderating effects. By this procedure, 302 of 17900 trials were deemed as outliers. Of the remaining trials, 2073 were incorrect responses and 1142 trials used no rotation.

Statistical analysis was performed as in the first study with linear mixed models using lme4 package (version 1.1-21; Bates, Mächler, et al., 2015) in R (version 3.5.1; R Core Team, 2018). Model parameters were estimated by maximum likelihood estimation using bobyqa algorithm wrapped by optimx package (version 2018-7.10; Nash & Varadhan, 2011) as optimizer. Model fit was calculated by using likelihood ratio tests to compare models with and without the fixed effect of interest. The resulting p-values were compared to a significance level of .05. For multiple comparisons of the same variables, the significance level was Bonferoni corrected. Visual inspection of residual plots did not reveal deviations from homoscedasticity or normality in any model.

Where possible, we report both the unstandardized effect sizes and confidence intervals calculated by using parametric bootstrapping with 1000 simulations in line with recommendations of Baguley (2009) and Pek and Flora (2018). While standardized effect sizes are routinely used for power analysis and meta analyses, unfortunately there does not exist an agreed upon way to compute standardized effect sizes in linear mixed models (Feingold, 2009; Hedges, 2007; Rights & Sterba, 2019). Nevertheless, linear mixed models offer several advantages over traditional use of ANOVAs. For example, linear mixed models allow simultaneous analysis of by-participant and by-item variances and thus eliminating the need to average over participants or items, while also allowing analysis of unbalanced data and achieving higher statistical power (Barr et al., 2013; Hilbert et al., 2019). These allowed the experiment to be controlled by time with participants finishing different numbers of trials and enabled the analysis of time as an independent variable.

Model building was based on the research of Barr et al. (2013) and Bates, Kliegl, et al. (2015), starting with a model with random intercepts and slopes for every appropriate fixed effect and reducing the model complexity by dropping non-significant variance components. Non-significant fixed effects were further stepwise removed from the model, such that effects which least decreased model fit were removed first and a model containing only significant fixed effects remained. Non-significant effects were then tested for an improvement of model fit by inclusion in the resulting model, while significant effects were tested for worsening of model fit by exclusion of the effect. Main effects for significant interactions were tested separately by splitting the interaction. The resulting models for each parameter are described in the results section.

As there is ongoing discussion about model selection based on theory or data or preferring complex instead of simpler models, we expect future research to cast doubt on the optimality of the currently suggested models. Although models without random slopes seem too anti-conservative (Barr et al., 2013) we report results of these simplest models for easier comparison.

5.3 Results

5.3.1 Descriptive Statistics

Means and standard deviations for both reaction time and accuracy are provided in table 10 for both side and degree as the main effects in question. Due to the time-controlled nature of the experiment, participants performed different number of trials. Average reaction time and accuracy are calculated for every participant. Mean and standard deviations are then reported between participants. Behavioral data for other conditions and summarized demographic data can be found at https://github.com/LeonardoJost/MRExperiment.

Table 10

Mean (SD) of Behavioral Data, Separated by the Side of the Congruent Base Figure (Columns) and Degree (Rows).

	Reaction time		Accuracy	
	Left	Right	Left	Right
0°	1824(643)	2074(578)	.97(.04)	.91(.12)
45°	2830(942)	2896(938)	.92(.09)	.92(.09)
90°	3664(1047)	3765(1151)	.87(.11)	.87(.12)
135°	4252(1477)	4343(1467)	.85(.13)	.86(.10)
180°	4804(1467)	4704(1525)	.81(.15)	.86(.11)

Note. Reaction time is reported in ms, accuracy in the proportion of correct answers. Mean reaction time is calculated for all correctly answered trials of each participant and then averaged over all participants. Mean accuracy is calculated for all trials of each participant and then averaged over all participants

5.3.2 Reaction Time

For the analysis of reaction time on the dataset of only rotated stimuli, model building resulted in a model with random intercepts and random slopes for degree and time (since start of the session) by participant and random intercepts by model. The interactions degree*time*side and degree*axis*side as well as the included two-way interactions and main effects, and the effects of previous experience were analyzed as fixed effects. Significant effects were found for experience, degree*time, degree*side, degree*axis, degree, and time. Reaction time improved with experience and time and increased with angular disparity. The effect of angular disparity was lower with increased time, for answers on the right side, and for rotations around the z-axis. (See table 11)

In a model without random slopes, the effects of experience ($\chi^2(1)=0.13$, $p=.719$) and degree*side ($\chi^2(1)=3.53$, $p=.060$) were not significant, but otherwise no differences in significance were found.

Table 11

Statistical Analysis of Reaction Time Including Axis on the Dataset of only Rotated Stimuli.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	2098.64	201.31			1686.14, 2487.78
Degree*Time	-671.04	98.53	$\chi^2(1)=46.28$	<.001	-863.22, -474.91
Degree*(Side=right)	-113.56	57.00	$\chi^2(1)=3.97$	.046	-223.75, -10.08
Degree*(Axis=Z)	-374.20	56.88	$\chi^2(1)=43.21$	<.001	-481.64, -272.71
Degree*(Time=0)* (Side=left)*(Axis=Y)	2062.09	114.75			1828.19, 2298.92
(Degree=0)*Time	-492.00	159.02			-782.20, -188.61
(Degree=0)*(Side=right)	149.81	63.55			32.34, 269.56
(Degree=0)*(Axis=Z)	387.88	63.46			269.76, 510.26
Experience=no	476.49	203.07	$\chi^2(1)=4.98$	.026	55.34, 862.66
Main Effects:					
Time	-1177.24	123.27	$\chi^2(1)=47.95$	<.001	-1417.51, -908.96
Side=right	34.65	26.41	$\chi^2(1)=1.72$	.190	-21.39, 90.13
Axis=Z	8.14	26.41	$\chi^2(1)=0.10$	.758	-41.39, 61.51
Degree	1457.13	93.63	$\chi^2(1)=78.83$	<.001	1272.79, 1648.52
Nonsignificant Effects:					
Degree*Side*Axis			$\chi^2(1)=0.11$	.746	
Degree*Side*Time			$\chi^2(1)=0.74$	.389	
Side*Axis			$\chi^2(1)=2.55$	.111	
Time*Side			$\chi^2(1)=0.24$	.626	

Note. The values for degree and time (since start of the session) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. The values for side=right, axis=z, and experience=no are all relative to the other respective condition. The intercept in this model represents the estimated reaction time in ms at time=0, degree=0, side=left, axis=y, experience=yes.

Table 12

Statistical Analysis of Reaction Time Excluding Axis on the Dataset of All Stimuli.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	2207.46	172.91			1870.43,2539.50
Degree*Time	-686.97	83.58	$\chi^2(1)=67.33$	<.001	-862.07, -523.55
Degree*(Side=right)	-149.63	48.20	$\chi^2(1)=9.63$	.002	-244.49, -53.33
Degree*(Time=0)* (Side=left)	1994.49	107.73			1787.55, 2192.85
(Degree=0)*Time	-473.26	141.90			-737.29, -195.83
(Degree=0)*(Side=right)	191.26	51.85			91.52, 291.70
Experience=no	323.01	158.03	$\chi^2(1)=3.89$	.049	1.76, 651.41
Main Effects:					
Time	-1125.96	117.44	$\chi^2(1)=48.18$	<.001	-1354.56, -881.59
Side=right	50.00	24.87	$\chi^2(1)=4.04$	.044	6.13, 101.95
Degree	1555.60	94.60	$\chi^2(1)=82.80$	<.001	1359.07, 1755.70
Nonsignificant Effects:					
Degree*Side*Time			$\chi^2(1)=0.32$	.570	
Time*Side			$\chi^2(1)=0.41$	.521	

Note. The values for degree and time (since start of the session) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. The values for side=right and experience=no are all relative to the other respective condition. The intercept in this model represents the estimated reaction time in ms at time=0, degree=0, side=left, experience=yes.

For the analysis of reaction time on the dataset of all stimuli, model building resulted in a model with random intercepts and random slopes for degree and time by participant and random intercepts by model. The interaction degree*time*side and the effects of previous experience were analyzed as fixed effects. Significant effects were found for experience, degree*time, degree*side, degree, time, and side. Compared to the model for only the rotated stimuli, reaction times were significantly slower for answers on the right side, but otherwise no differences in significance were found. (See table 12)

In a model without random slopes, the effects of experience ($\chi^2(1)=0.11$, $p=.740$) and side ($\chi^2(1)=3.65$, $p=.056$) were not significant, but otherwise no differences in significance were found.

5.3.3 Accuracy

Accuracy was analyzed by a general linear mixed model, which used a binomial distribution. For the analysis on the dataset of only rotated stimuli, model building resulted in a model with random intercepts and random slopes for degree, time (since start of the session), and axis by participant and random intercepts and random slopes for time by model. The interactions degree*time*side and degree*axis*side and the effects of previous experience were analyzed as fixed effects. Significant effects were found for degree*axis, degree, and time. Accuracy was significantly higher with increased time and significantly lower with increased angular disparity. The effect of degree was lower for rotations around the z-axis. (See table 13)

In a model without random slopes, the effects of degree*time ($\chi^2(1)=5.12$, $p=.024$), degree*side ($\chi^2(1)=3.94$, $p=.047$), and axis ($\chi^2(1)=7.41$, $p=.006$) were significant, but otherwise no differences in significance were found.

For the analysis of accuracy on the dataset of all stimuli, model building resulted in a model with random intercepts and random slopes for degree and time by participant and random intercepts and random slopes for time by model. The interaction degree*time*side and the effects of previous experience were analyzed as fixed effects. Significant effects were found for degree*side, degree, and time. Compared to the model for only the rotated stimuli the decrease of accuracy with increasing angular disparity was significantly smaller for answers on the right side, but otherwise no differences in significance were found. (See table 14)

In a model without random slopes, the effect of degree*time ($\chi^2(1)=7.39$, $p=.007$) was significant and the effect of time ($\chi^2(1)=3.58$, $p=.058$) was not significant, but otherwise no differences in significance were found.

Table 13

Statistical Analysis of Accuracy Including Axis on the Dataset of Only Rotated Stimuli Using a Binomial Distribution.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	3.14	0.25			2.67, 3.65
Degree*(Axis=Z)	0.63	0.11	$\chi^2(1)=31.38$	<.001	0.41, 0.86
Degree*(Axis=Y)	-1.05	0.11			-1.28, -0.84
(Degree=0)*(Axis=Z)	-0.60	0.15			-0.92, -0.32
Time	0.34	0.16	$\chi^2(1)=4.52$	.033	0.02, 0.66
Main Effects:					
Axis=Z	0.12	0.08	$\chi^2(1)=2.06$	.151	-0.04, 0.29
Degree	-0.75	0.09	$\chi^2(1)=45.61$	<.001	-0.93, -0.58
Nonsignificant Effects:					
Side=right	0.10	0.05	$\chi^2(1)=3.55$	.059	-0.00, 0.20
Degree*Side*Axis			$\chi^2(1)=0.57$	.450	
Side*Axis			$\chi^2(1)=0.09$	.761	
Degree*Side*Time			$\chi^2(1)=0.00$	.955	
Degree*Time			$\chi^2(1)=0.93$	.335	
Degree*Side			$\chi^2(1)=3.83$	.050	
Time*Side			$\chi^2(1)=0.88$	.348	
Experience=no	-0.48	0.29	$\chi^2(1)=2.71$	.100	-1.03, 0.07

Note. The values for degree and time (since start of the session) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. The values for side=right, axis=z, and experience=no are all relative to the other respective condition. The intercept in this model represents the estimate at time=0, degree=0, axis=y.

Table 14

Statistical Analysis of Accuracy Excluding Axis on the Dataset of All Stimuli Using a Binomial Distribution.

Variable	Estimate	SE	Test statistic	p	95% CI
Intercept	2.99	0.21			2.60, 3.46
Degree*(Side=right)	0.39	0.10	$\chi^2(1)=15.76$	<.001	0.21, 0.61
Degree*(Side=left)	-0.93	0.09			-1.12, -0.74
(Degree=0)*(Side=right)	-0.37	0.11			-0.62, -0.16
Time	0.36	0.16	$\chi^2(1)=4.74$	.030	0.03, 0.70
Main Effects:					
Side=right	0.05	0.05	$\chi^2(1)=1.19$	.276	-0.04, 0.15
Degree	-0.74	0.07	$\chi^2(1)=54.40$	<.001	-0.89, -0.59
Nonsignificant Effects:					
Degree*Side*Time			$\chi^2(1)=0.02$	.879	
Degree*Time			$\chi^2(1)=2.77$	.096	
Time*Side			$\chi^2(1)=0.78$	.378	
Experience=no	-0.40	0.27	$\chi^2(1)=2.04$	.154	-0.92, 0.18

Note. The values for degree and time (since start of the session) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. The values for side=right and experience=no are all relative to the other respective condition. The intercept in this model represents the estimate at time=0, degree=0, side=left.

As we observed nonlinear behavior in the accuracy data averaged by participant, we performed separate comparisons by degree for both rotation axis and the correct side. For this, we used a model with random intercepts and random slopes for time by both participant and model. As fixed effects, we included both axis and side for all angles for comparison purposes. At 0° answers on the right side were significantly less accurate than answers on the left side, while at 180° they were significantly more accurate. No differences were found at 45°, 90°, and 135°. Rotations around the z-axis produced significantly more accurate answers at 135° and 180° compared to rotations around the y-axis. (See table 15)

Table 15

Statistical Analysis of Accuracy Separated by Angle.

Variable	Estimate	SE	Test statistic	p	95% CI
0°					
Intercept	4.31	0.38			3.73, 5.35
Side=right	-1.40	0.31	$\chi^2(1)=22.67$	<.001	-2.18, -0.79
45°					
Intercept	3.06	0.22			2.65, 3.48
Side=right	0.01	0.11	$\chi^2(1)=0.01$	.933	-0.21, 0.24
Axis=Z	-0.15	0.11	$\chi^2(1)=1.72$	.190	-0.39, 0.08
90°					
Intercept	2.33	0.21			1.95, 2.76
Side=right	0.04	0.09	$\chi^2(1)=0.15$	.701	-0.14, 0.21
Axis=Z	-0.19	0.09	$\chi^2(1)=3.98$	.046	-0.38, -0.00
135°					
Intercept	1.90	0.19			1.53, 2.26
Side=right	0.03	0.09	$\chi^2(1)=0.11$	.739	-0.14, 0.21
Axis=Z	0.35	0.09	$\chi^2(1)=15.59$	<.001	0.17, 0.53
180°					
Intercept	1.42	0.20			1.06, 1.85
Side=right	0.43	0.12	$\chi^2(1)=12.48$	<.001	0.21, 0.65
Axis=Z	0.63	0.12	$\chi^2(1)=27.01$	<.001	0.40, 0.88

Note. Intercepts in this model represent the estimate for axis=y, side=left.

5.3.4 Linearity of Data

As we observed nonlinear behavior in the data, we analyzed the influence of additional polynomial functions of angular disparity on both reaction time and accuracy. In these cases, a significant improvement of model fit was found for quadratic and cubic polynomials. All estimated polynomials were monotonously increasing for reaction time and monotonously decreasing for accuracy in the value range of interest. (See table 16)

Table 16

Estimated Coefficients of Higher Order Polynomials for the Relationship of Reaction Time and Accuracy With Angular Disparity at Units of 100°.

Dataset	Degree	x^1	x^2	x^3	x^4	Test statistic	p
Reaction time	1	2062.09					
including axis	2	2922.18	-407.6			$\chi^2(1)=35.85$	<.001
	3	4948.50	-2447.06	611.08		$\chi^2(1)=8.66$	.003
Reaction time	1	1994.49					
excluding axis	2	2689.69	-371.33			$\chi^2(1)=68.09$	<.001
	3	2757.85	-464.58	33.41		$\chi^2(1)=0.17$	.682
	4	1863.39	2111.76	-2272.16	637.82	$\chi^2(1)=9.28$	.002
Accuracy	1	-1.05					
including axis	2	-2.57	0.66			$\chi^2(1)=25.29$	<.001
	3	-5.60	3.65	-0.88		$\chi^2(1)=5.19$	.023
Accuracy	1	-0.93					
excluding axis	2	-1.93	0.47			$\chi^2(1)=23.00$	<.001
	3	-1.80	0.30	0.06		$\chi^2(1)=0.10$	.755
	4	0.42	-5.02	4.47	-1.17	$\chi^2(1)=3.80$	.051

Note. Degree describes the degree of the fitted polynomial and x^1-x^4 describe the coefficients. Test statistics were computed in comparison to the polynomial of one lower degree.

5.3.5 Stimulus Library

Due to the work of Peters and Battista (2008) a stimulus library of cube figures is already available. In an attempt to expand the applications and research possibilities, we have written an extended, more flexible generation process in R (R Core Team, 2018). To keep comparisons with older work valid, we have used the same cube structure for the base figures as Peters and Battista (2008), although we do not exactly reproduce the rotated versions. In the following, we describe the reasoned small changes we have made as well as the extensions, applications, and availability.

5.3.5.1 Differences

As the rotation around three axes is not commutative and the stimuli are typically generated by a stepwise rotation around the three axes, the order of applying the rotation matters. As can be seen from the stimuli of the Peters and Battista (2008) library, the order of rotation is x-axis, then y-axis, then z-axis. This produces occlusions at rotation angles of 90° and 270° around the y-axis. In these cases, the x-axis and the previously performed rotation around it is transformed into the z-axis and thus, both produce a rotation around the z-axis of the final figure. The resulting figure then shows no rotation around the x-axis and as such no rotation in depth perpendicular to the z-axis. As a result, the faces of the cubes in the direction of the z-axis facing the top and bottom are not visible.

We have chosen a rotation order of x-axis, then z-axis, then y-axis. While in the same case of a rotation around the z-axis of 90° or 270° the other two axes are transformed into each other, occlusions occur at these angles independent of other rotation angles and orders. Additionally, this order keeps the rotation around the y-axis as a picture-plane rotation.

Compared to the library of Peters and Battista (2008) the rotation around the x-axis is also in the other direction, but this presents only a difference in naming the stimuli.

5.3.5.2 Properties, Extensions, and Applications

5.3.5.2.1 Possibility of Different Base Angles

As the rotations around different axes in the library of Peters and Battista (2008) are askew, that is, no identical figure appears for all three rotation directions, we have implemented the usage of any possible base angle.

5.3.5.2.2 Possibility of Rotation and Mirroring

Figures can be mirrored across all natural planes (see figure 11). While these can also be constructed by rotation of one mirror image, this can be used to test effects of orientation more easily. Moreover, a rotation around all natural axes is possible at all possible angles, not only around single axes, but also around all axes at once.

5.3.5.2.3 Colors and Background

We have implemented the possibility to use different colors for both the figure and the background. For coloring the figure, this includes both the color of the borders as well as the color of the faces, allowing also different checkered patterns. For the background, this allows a plain background in any color as well as using other images as background.

5.3.5.2.4 Occlusions

Occlusion occurs when rotation angles around either x- or z-axis are multiples of 90°. For each axis the number of visible faces is reduced by one. Occlusions can be avoided by choosing a base rotation angle around the x- and z-axis.

5.3.5.2.5 Centering

As figures are typically not centered in the image (one can see in figure 12 that the horizontal parts of the figures are aligned vertically, but as the vertical arms of the figure differ in length, one figures appears higher than the other), we have implemented two options for centering figures. The first option, centering by weight, distributes equal weight to the individual cubes and centers the figure, such that the center of mass lies in the origin. The second option, optical centering, centers the bounding box of the figure, such that the expansion in the three natural directions is centered. Optical centering allows the representation of the largest figures at given image dimensions.

5.3.5.2.6 Extension Beyond Cube Figures

Not only can custom cube models be constructed, and the process is easily adaptable for other geometric shapes with polygonal surfaces, but an additional extension is possible by building models from many small cubes. As cube size converges to pixel size, three-dimensional rotations of arbitrary shapes are possible (see figure 13), although at a greater cost of computing power. This can be used for in-depth rotations of real-world objects or human figures in a comparable way to cube figures, if they are described by a three-dimensional point matrix.

5.3.5.2.7 Image Formats

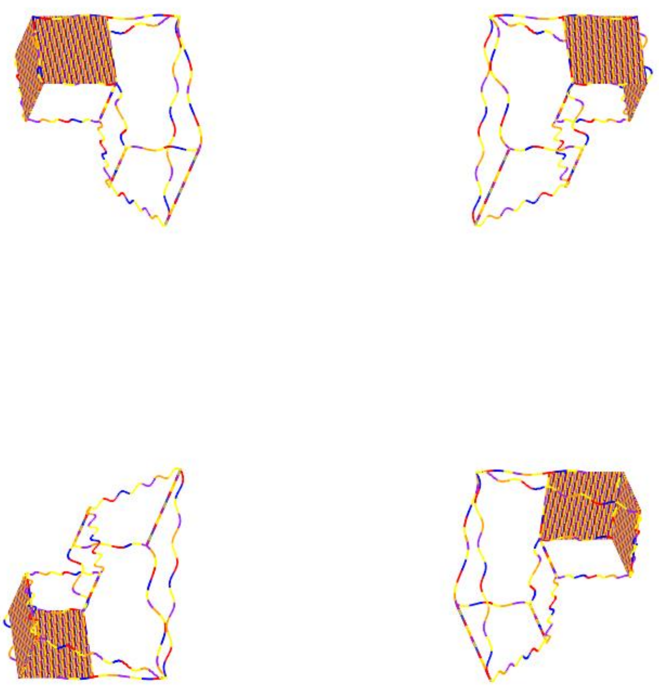
In- and output is read and written using magick package (version 2.0; Ooms, 2018), which supports multiple image formats, including most common and lossless formats.

5.3.5.2.8 Availability

We make both pregenerated cube figures and the code for generation of individual figures available at https://github.com/LeonardoJost/MRlibrary. Upon usage, we advise to state all values of the parameters given in table 9.

Figure 13

Depictions of an Arbitrary Figure, Employing Both Varying Colors and Curved Lines as Well as Non-Rectangular Faces, and Mirrorings and Rotations of the Figure.



Note. Top left: Base model. Top right: Mirrored across the yz-plane. Bottom left: Rotation around the x-axis by 180°. Bottom right: Rotation around the z-axis by 180°.

5.4 Discussion

Our study provides insight that the newly developed test is applicable for measuring mental rotation ability. With this test all stimuli can be analyzed, although small differences between the sides of the correct base figure persist. We have found some minor differences in reaction time but some larger differences in accuracy. While the behavior of reaction time and accuracy by degree shows changes by side and with different rotation axes, the overall results are comparable to the behavioral data of "same" stimuli in the design of R.N. Shepard and Metzler (1971) with two images regarding effects of degree. That is, for both sides of the congruent base figure we observe the behavior expected of "same" stimuli. Moreover, if only one base figure were used to solve the task, we should have found reaction time differences comparable to those between "same" and "different" trials in two-figure mental rotation tasks. As this is not the case, we assume that the unique 0°-position is used by participants and the same mental rotation process is measured in all trials.

Regarding reaction time, we have found the expected relationship to degree and time. Reaction time increases with larger angles and decreases over time, with larger improvements at larger angles. We have found differences in the interaction of degree and side, but on both sides our results show the expected increase of reaction time at larger angles. Moreover, the differences by side are smaller than by axis, which are typically assumed to be the same. Possibly of the same magnitude are effects of rotation direction, which were not analyzed and are also typically grouped together. Additionally, the improvement over time at all angles did not differ by side, which leads us to the assumption, that the same mental transformation processes are used on both sides. This is also supported by the lower reaction times of participants with previous experience with the two-figure layout, although learning effects are not fully understood (Heil et al., 1998; Meneghetti et al., 2017; Rahe, Ruthsatz, Jansen, et al., 2019; Uttal et al., 2013).

For accuracy we have found differences by side for angular disparities of 0° and 180° as well as differences by axis. Compared to reaction time, accuracy shows no significant differences for previous experience and for the interaction of degree and time, but results point in the same direction as for reaction time. In the case of side and axis, the results indicate higher accuracies at lower reaction times and cannot be justified by speed-accuracy trade-offs. The differences in accuracy at 0° are quite surprising, as no mental rotation is necessary, yet average accuracy was only at 91% for answers on the right side (compared to 97% for the left side). One possible explanation might be the left to right reading direction and thus figures were possibly also compared in that order, although reaction time differences would be better explained by this than accuracy differences. Also, we have observed a quite unexpected increase in accuracy for rotations around the z-axis at larger angles. While this might be influenced by the layout of the figures or the base angles of the figures, an interaction with the side of answer was not significant. A possible influence can be seen in figure 12 for one model, but the other models show a similar position of the characteristic arm. For a rotation around the z-axis, the characteristic arm of the figure stays close to the two base figures, whereas for a rotation around the y-axis the optical distance between stimulus and base figures increases. By the same matter, there is a possible overlap of the figure with the fixation cross between trials for rotations around the z-axis. Another difference can be seen in the base angles, which produce different directions of the front- and top-facing arms, which would point straight forward or upwards if no base angles were used. As these explanations are speculative, it remains unclear, how and why accuracy differences by axis occur. Further research needs to be done on reproducibility and explanations of these phenomena and accuracy should be interpreted with more caution.

Similarly as for reaction time, overall accuracy patterns show smaller differences by side than by axis and no interactions with time, which can be interpreted in support of the notion that the same mental transformations are performed on both sides as in the two-figure case.

Average reaction times are also similar to other studies using comparable stimuli but different designs. For example, in the first study we used two main sessions of 30 minutes duration on different days and found on average slightly lower reaction times. Meneghetti et al. (2017) used multiple training sessions of approximately 35 minutes duration and found similar reaction time in their second session. Thus, our layout enables the completion of more analyzable tasks in time and achieves the desired increase in power.

5.4.1 Nonlinearity

Although residual plots for models using linear dependency of reaction time on degree show no visual deviations, significant higher order polynomial relationships were observed. Whereas the original paper of R.N. Shepard and Metzler (1971) showed perfect linearity in reaction time, this is not unanimously reproduced (e.g. Jolicœur et al., 1985), and, although often not discussed, inspection of the data in other work rarely shows perfect linearity. Thus, we believe that nonlinearity is not a major concern in discussing if our setup can be used to analyze the same mental processes as the original design. While the linearity allows a simple theoretical framework, namely a fixed mental rotation speed, we have no simple explanation for the nonlinearity. Differing speed-accuracy trade-offs might be one reason, but different strategies analyzing mirrorings instead of rotation might also contribute. In our opinion, the important findings are that reaction time increases and accuracy decreases with increasing angular disparity and deviations from linearity do not show large effects. Until further research is available, we believe, if nonlinearity can be observed in the data, further investigating interactions should be warranted even if they are not significant, as for example in our data in the case of accuracy differences by side for angular disparity of 180°.

5.5 Limitations

One limitation is the inference of mental rotation only from behavioral data. Although it seems plausible, we cannot know if strategies incorporate the usage of both stimuli or if stimuli are actually rotated in the mind from this behavioral data alone and thus cannot infer that the same strategies are used in our design as in the two-figure case. However, the investigation of strategies was not the main goal of our study. Moreover, effects of increasing mental fatigue in time as well as effects of the shape and layout of the cube figures could have influenced our results.

5.6 Conclusion

In conclusion, we have validated a new layout for chronometric mental rotation tests utilizing three figures. While small differences in performance need to be further analyzed, overall, our design is appropriate for analyzing mental rotation performance. The new layout offers an increase of power while also solving possible problems with analysis of accuracy. Our theoretical framework of layout parameters as well as the described stimulus library provide possibilities for future research and implementation.

Using linear mixed models for statistical analysis allowed our work to be one of the first to analyze the progress of participants during the experiment. While we analyzed the absolute improvement in ms and the proportion of correct answers with more time spent on the task, future research could also analyze improvements relative to participants mean performance or with the number of items solved. We also found performance improvements for participants who had experience with mental rotation experiments. There is potential for future work to specify the variable experience more precisely and further explore the relationship to performance.

6 Third Study: Learning Processes Within Sessions and Treatment Effects[4]

The first two studies have shown an improvement of mental rotation performance within the sessions. In the second study, these improvements were also larger for larger angular disparities. An interesting question arising from this is how improvements within sessions affect the detection and interpretation of improvements between sessions due to treatments. The usefulness and relevance of such an approach was explored in this third study.

6.1 Goal and Hypotheses

Linear mixed models have shown statistical advantages over ANOVAs. One advantage is the possibility to analyze the progress during the time course of an experiment. Here, we demonstrate how accounting for time within sessions in pre-posttest designs can improve control over practice effects and aid the interpretation of results. In traditional analyses, a treatment effect is combined with (unwanted) practice effects in a treatment group. This is compared to a control group, which should only be influenced by practice effects. However, the isolation of practice effects in control groups is still not optimal and the variance of practice effects between participants can impact the detection of treatment effects. Moreover, when performance differences between groups are detected, it is clear that the interpretability of posttest differences is impacted. The same should be clear for differences in practice effects between groups, which could in theory occur equally often but are rarely investigated. We present an approach, which through inclusion and manipulation of time within sessions estimates practice effects within sessions and has the following advantages:

First, it allows better detection of the magnitude and change of practice effects, which can in themselves be interesting.

[4] The results presented in this chapter are under review: Jost, L. & Jansen, P. (under review). Using linear mixed models to analyze learning processes within sessions improves detection of treatment effects: An exemplary study of chronometric mental rotation.

Second, by the separating practice and treatment effects, it allows better estimation of the "true" treatment effect on test performance.

For this, we have performed multiple analyses on the dataset of the second study[5]. For mental rotation, improvements by repetition have been shown within and between multiple sessions. These practice effects are an unwanted influence in repeated testing designs and are observable for many other cognitive tests (Calamia et al., 2012; Goldberg et al., 2015).

We have split the data to represent multiple testing sessions or a possible treatment session consisting of the repetition of the task. In the following analyses of these sessions, we investigate the detection of a treatment effect compared to no treatment between sessions or to correctly identify the null effect if no treatment was used. We compare an analysis using the suggested inclusion and manipulation of time within sessions to a more traditional analysis using only the testing sessions and hypothesize that the inclusion of time will improve the detection.

In the following, we will first introduce the general principle and show the possibilities of analyzing learning within one group and multiple testing sessions. In a second step, we will simulate two groups in a pre-post design and show how the analysis of time can aid interpretation of treatment effects even where traditional analyses might fail. We provide some constructed cases to demonstrate potential problems as well as simulations regarding their occurrences in practice.

6.2 Method

6.2.1 Description of the Dataset

We have reused the dataset of the second study. We focus only on the analysis of reaction time as it shows easier interpretation and larger effects than accuracy and is typically the main variable of interest in mental rotation tasks. The dataset contains a total of 15525 observations of 41 participants and 16 different stimuli (item types) for the analysis of reaction time. For each

[5] The data is openly available at https://osf.io/dr9mv/ (DOI 10.17605/OSF.IO/DR9MV).

observation, the time of the answer since the start of the session (time) is reported. Next to the main variables of interest, the dataset also includes other variables (degree, side, experience), which are included as covariates but are not of primary interest here.

To simulate the effects of repeated testing, this dataset was split into three blocks of 10 minutes by the time of the answer of the participants. These blocks will be analyzed as simulated pre- or posttests or as a training condition employing the repetition of the task. As a result, for comparison between adjacent blocks we expect no improvement. This comparison will be used either as a control group or as a treatment without effect. For comparison between the first and third block, we expect to find an improvement by the repetition of the task in the second block, that is, the treatment.

6.2.2 Statistical Analysis

Statistical analysis was performed similarly to the second study with linear mixed models using lme4 package (version 1.1-23; Bates et al., 2015) in R (version 4.0.3; R Core Team, 2018). Model parameters were estimated by maximum likelihood estimation using bobyqa algorithm wrapped by optimx package (version 2020-4.2; Nash & Varadhan, 2011) as optimizer. P-values were obtained by using likelihood ratio tests to test for improvement of model fit by inclusion of the fixed effect of interest and compared to a significance level of .05.

We do not report effect sizes because the relevant effect sizes for the interpretation have already been reported in the second study. For the application of the ideas presented here, the detection of effects is the deciding factor not the actual size of the effect. Although the existence of effects differs from significance (Amrhein et al., 2019) we will use significance to judge whether an effect will be found or how interpretable effects are.

The data contains two random effects: the participants and the item type and we used random intercepts and random slopes for degree and time by participant and random intercepts by item based on the suggestion of Matuschek et al. (2017). For the analysis of only blocks, we replaced the random slope by time by a random slope by block. Moreover, we included the main effect of experience and the interaction between degree and side as these proved to be significant and possible covariates in the original dataset. For comparison purposes, these were included in all analyses regardless of their significance.

To compare the effects of time, we performed two analyses for each dataset. In one analysis, we excluded all effects of time and analyzed the interaction between degree and block. In the other analysis, we included the effects of time and analyzed the interaction between degree, block, and time. Degree was included as it is the main moderator of difficulty in mental rotation tasks with larger improvements for larger angles. For the comparison of different groups, we analyzed the additional interaction with group.

Starting from these models non-significant fixed effects were stepwise removed from the model, such that effects, which least decreased model fit were removed first and a model containing only significant fixed effects remained. This was performed only for the examples but not for the random simulations. The analysis of main effects contained in significant interactions was performed according to Levy (2014). Degree was centered such that main effects show the average improvement over all angles. Time was normalized such that time 0 was set to the end of the first block and main effects of block would indicate additional improvements between the first and the next block that could not be explained by improvements over time.

For the visualization of the data and the changes within sessions, we use generalized additive models as the possibly best representation of true effects. These were implemented using ggplot2 package (Wickham, 2016) in R (R Core Team, 2018). A visualization using linear improvements or block-wise approximation would be biased for one of the approaches discussed here.

6.3 Results

6.3.1 Analyses of Performance Within One Group

6.3.1.1 Analysis of All Blocks Without Treatment

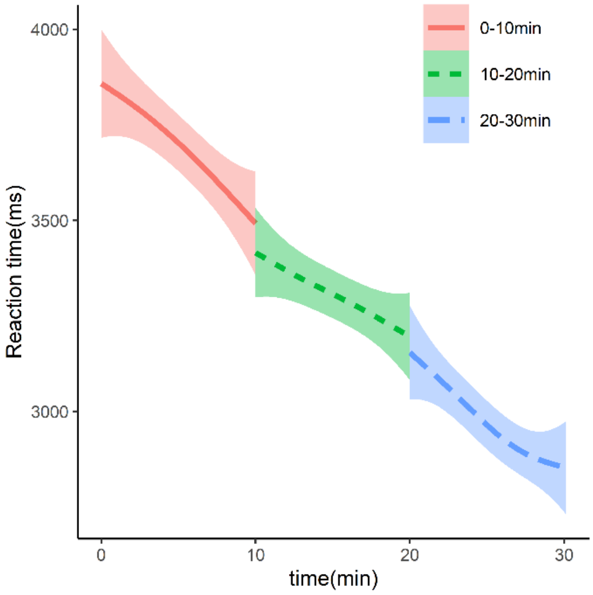
First, we start with the full dataset. The overall improvement over time during the separate blocks is shown in figure 14.

For the analysis of only blocks, there is a significant improvement between blocks ($\chi^2(2)=48.77$, $p<.001$). With the inclusion of time, the main effect of block are not significant anymore ($\chi^2(2)=1.97$, $p=.373$). The improvement over time can explain all improvements over the blocks but, in this case, does not vary significantly between blocks ($\chi^2(2)=0.33$, $p=.848$). As a result, the improvement over time can also interpolate the improvement during the second block.

Note that traditionally, one could include time as a covariate by centering time within each block. This reduces the variance within blocks by accounting for these improvements. However, this effect is quite small in the present data and does not change the significance of the results, as still the average performance within each block is compared. As the effect of time does not vary between blocks, the choice of centering only influences the intercept. However, in the case that the effect of time differs between blocks, the choice of time 0 needs further consideration. In this case, the main effect of block describes the difference between blocks at time 0 (Levy, 2014). Thus, time was normalized such that time 0 was set to the end of the first block and main effects of block would indicate additional improvements between the first and the next block that could not be explained by improvements over time.

Figure 14

Changes of Reaction Time Over the Full Session, Separated Into Three Blocks.



6.3.1.2 Analysis of a Treatment Between Two Blocks

Now, we look at only the first and the third block, where the second block serves as the treatment. We remove the time gap between the blocks and move the third block 10 minutes forward in time. For the analysis of only blocks, there is a significant improvement by block ($\chi^2(1)=49.47$, $p<.001$). With the inclusion of time, the main effect of block remains significant ($\chi^2(1)=46.95$, $p<.001$). This shows that the repetition of the task in the second block causes an improvement, which can no longer be explained if the time of the second block is not accounted for. The effect of block in the time analysis now describes the difference between the end of the first block and the start of the third block. That is, when compared to a hypothetical situation without a second block there is a significant improvement, that is, a treatment effect. This type of analysis will thus be used to evaluate the effectiveness of treatments.

6.3.2 Analysis of Treatment Effects Between Groups in a Pre-Post-Design

Now, we will assign participants to groups and compare the detection of treatment effects. We will start by demonstrating the general principle and follow with special cases, where the groups differ in performance (as is often observed in real studies) and in their improvements within sessions (which seems equally probable but is rarely measured).

6.3.2.1 Analysis of a Treatment Between Groups

We start with a case, where treatment and control group show a comparable pretest performance. For this, we duplicated the data and used one set for the treatment group and the other set for the control group. Both groups use the first block as pretest. The control group uses the second block as posttest and the treatment group uses the third block as posttest (thus including the second block as treatment). Again, block three is moved forward in time by 10 minutes. Not surprisingly, both types of analyses identify a treatment effect in the form of a significant interaction of block and group ($\chi^2(1)=12.36$, $p<.001$ and $\chi^2(1)=22.94$, $p<.001$). In such a case, on would typically perform separate analyses of each group. Here, these reiterate the results of the previous two analyses. Using only blocks for the analysis there is a smaller improvement in the control group while including time shows no significant improvement in the control group by block. Both analyses reveal that the treatment indeed improves performance.

Figure 15

Changes of Reaction Time Within Sessions, Separated by Groups. Groups Differ in Pretest Performance With Better Performers in the Treatment Group. Pre- and Posttest Are Separated by Time 0.

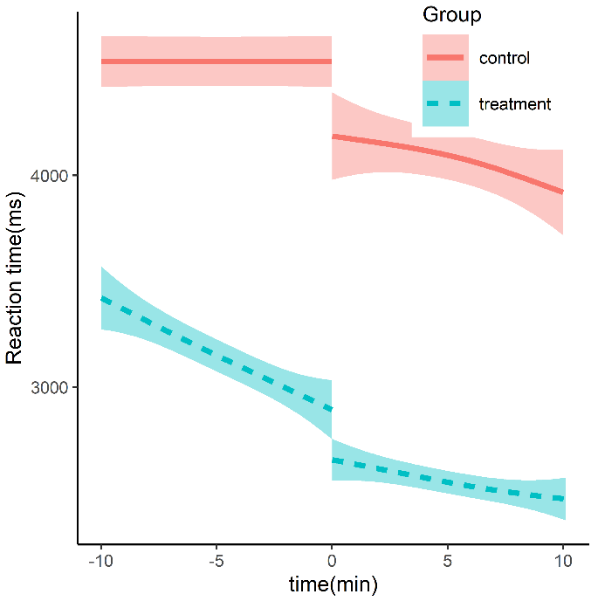
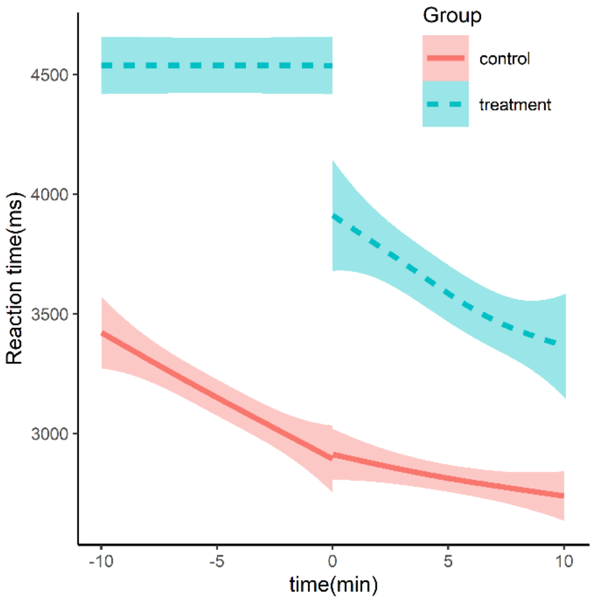


Figure 16

Changes of Reaction Time Within Sessions, Separated by Groups. Groups Differ in Pretest Performance With Better Performers in the Control Group. Pre- and Posttest Are Separated by Time 0.



6.3.2.2 Analysis of a Treatment and Between-Groups Differences in Pretest Performance

Now, we assign groups to participants based on their performance in the first block (pretest). Typically, treatment effects are harder to interpret in these cases. On the one hand, one expects a larger improvement in the treatment group compared to the control group. On the other hand, one would expect smaller improvements for better performers.

6.3.2.2.1 Better Performers in the Treatment Group

All participants with faster than median reaction time are assigned to the treatment group while the slower participants are in the control group. Both groups use the first block as pretest. As before, the control group uses the second block as posttest and the treatment group uses the third block as posttest. Block three is moved forward in time by 10 minutes. The improvement over time for this dataset is shown in figure 15.

For both analyses, the interaction between block and group is not significant ($\chi^2(1)=1.42$, p=.234 and $\chi^2(1)=0.08$, p=.778). As expected, the improvement is smaller for better performers and this cancels out the treatment effect. The analysis of time shows a significant main effect of block and a significant interaction of time, block, and group. These could be investigated further to identify possible reasons for the null effect of the treatment. However, both analyses should reach the same conclusion: More research is necessary in such a case.

6.3.2.2.2 Better Performers in the Control Group

This dataset resembles the previous one, but the slower participants are now in the treatment group and the faster participants are in the control group (see figure 16). Both analyses show significant interactions of block and group with larger improvements for the treatment group ($\chi^2(1)=13.228$, p<.001 and $\chi^2(1)=37.75$, p<.001). In the analysis of the control group separately, the inclusion of time can explain the improvement of the control group ($\chi^2(1)=0.03$, p=.864), whereas the analysis without time also shows significant improvement in the control group ($\chi^2(1)=20.49$, p<.001). This indicates that the improvement in the control group is only due to a practice effect and indicates a treatment effect. However, again, in such a case more research should be conducted.

6.3.2.3 Analysis of a Treatment and Between-Groups Differences in Practice Effects

Before, we separated the participants by their performance. Now, we will separate them by their improvement from the first to the second block, that is, their practice effect. These are very constructed worst cases for the analysis of only blocks because it cannot distinguish between practice and treatment effects.

We will start with assigning those participants to the control group, which show the largest practice effects (see figure 17). In this case, the analysis without time shows a significant interaction of block and group ($\chi^2(1)=4.66$, p=.031). The coefficients however lead to the wrong conclusion: They reveal a larger improvement in the control group. The analysis including time on the other hand can correctly identify the larger improvement for the treatment group ($\chi^2(1)=4.17$, p=.041).

In another example, we assign the participants with larger practice effects to the treatment group but without an actual treatment. That is, we use the first two blocks for both the control group and the practice group and there should be no treatment effect (see figure 18). The analysis without time nevertheless shows a significantly larger improvement for the treatment group ($\chi^2(1)=30.29$, p<.001), which can be explained with the inclusion of time ($\chi^2(1)=0.12$, p=.732).

Figure 17

Changes of Reaction Time Within Sessions, Separated by Groups. Groups Differ in Improvement With Better Learners in the Control Group. Pre- and Posttest Are Separated by Time 0.

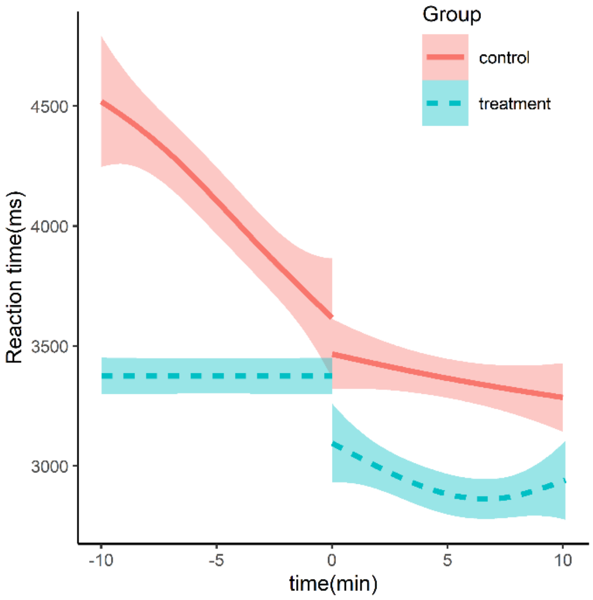
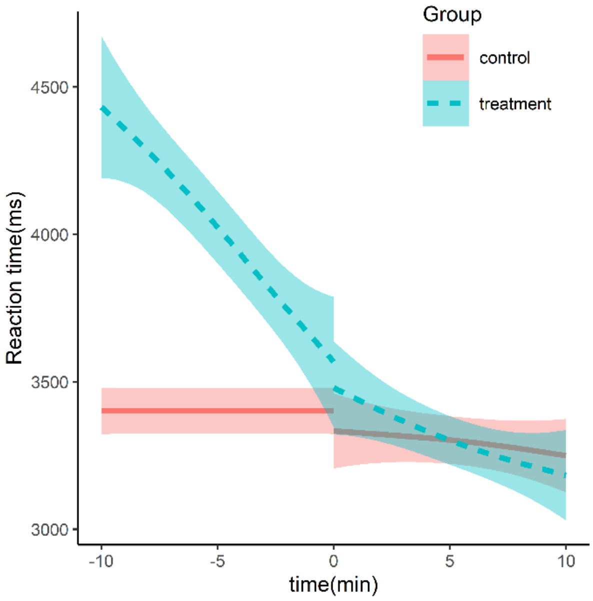


Figure 18

Changes of Reaction Time Within Sessions, Separated by Groups. Groups Differ in Improvement With Better Learners Labeled as Treatment Group but Without an Actual Treatment. Pre- and Posttest Are Separated by Time 0.



6.3.2.4 Random Group Allocations

The previous cases were all constructed extreme cases. To show possible occurrences in reality, we have performed two times 1000 random simulations of group allocations. One set of simulations used the second block as treatment (block one and three as pre- and posttest) and the other set used no treatment (block one and two as pre- and posttest). For each of the simulations, 20 or 21 of the 41 participants were randomly selected and assigned to the treatment group. The other participants were assigned to the control group. For every simulation, we compared the detection of a significant treatment effect between groups for both discussed types of analyses. In cases of convergence issues in any analysis, a new random allocation was generated. A summary of the significance of p-values is shown in table 17. For comparison, we also included an analysis employing the change over time within sessions as a covariate by centering within each block. This analysis however only differed marginally from the analysis using only the blocks (differing in significance for only six out of the 2000 simulations). This shows that the differences in the results of the proposed approach are not simply due to the inclusion of time but also the choice of the manipulation of time between blocks.

Table 17

*Comparison of the Significance of the Block*Group Interaction in 1000 Random Simulations Using the Treatment and 1000 Random Simulations Without a Treatment for the Analyses Accounting for Time (Rows) and not Using Time (Columns).*

Time\no time	Treatment		No treatment	
	p<.05	p>.05	p<.05	p>.05
p<.05	641	131	3	51
p>.05	138	90	54	895

Figure 19

Changes of Reaction Time Within Sessions for the Random Group Allocation Using a Treatment, Where at Least One Analysis Does not Show a Significant Treatment Effect. Pre- and Posttest Are Separated by Time 0. Left: Analysis With Time Does not Show a Significant Treatment Effect. Center: Analysis Without Time Does not Show a Significant Treatment Effect. Right: Neither Analysis Shows a Significant Treatment Effect.

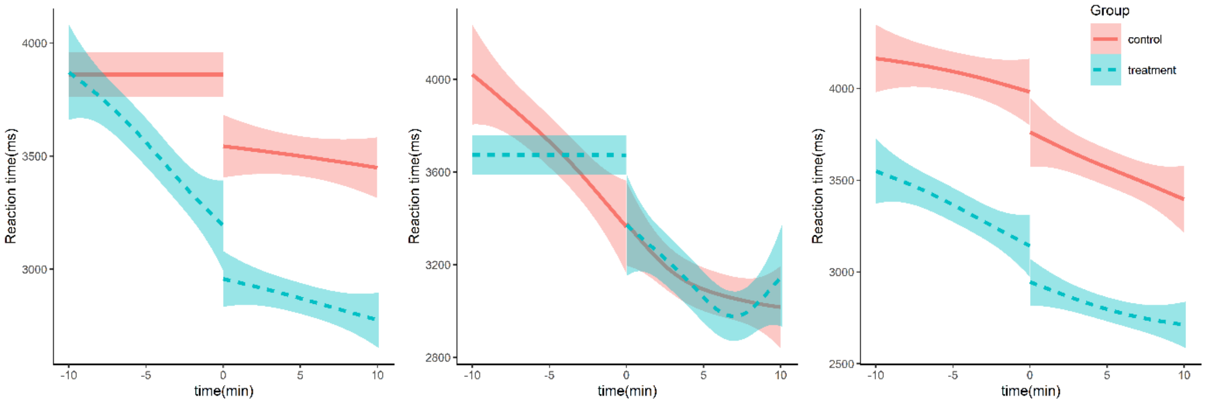
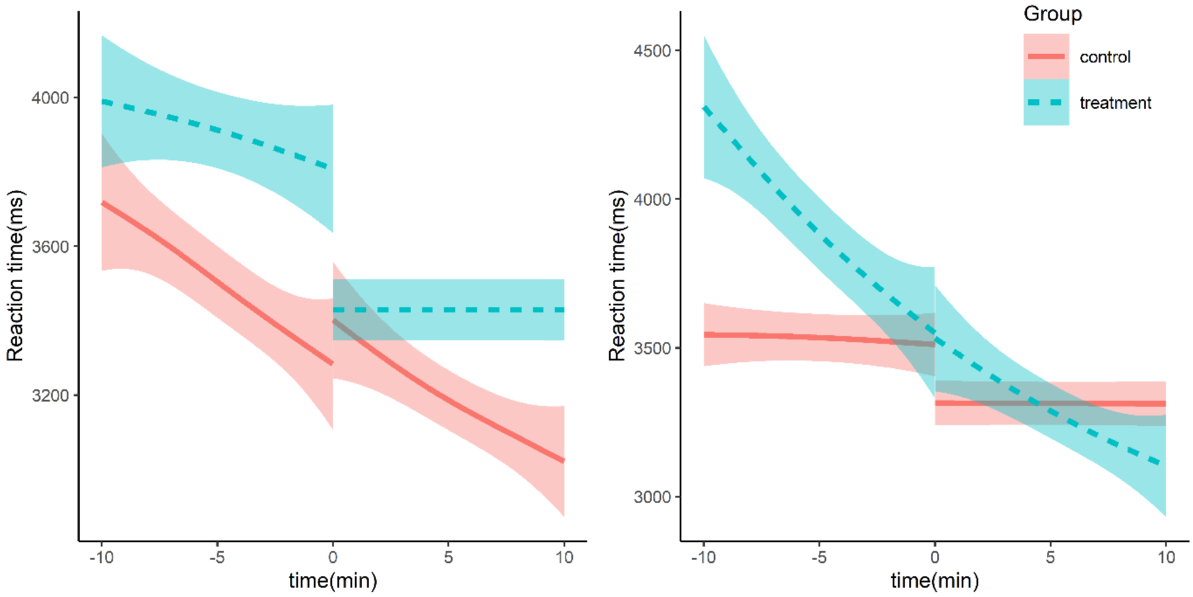


Figure 20

Smoothed Conditional Means of Reaction Time for the Random Group Allocation Using no Treatment, Where at Least One Analysis Shows a Significant Treatment Effect. Blocks Are Separated by Time. Left: Analysis With Time Shows a Significant Treatment Effect. Right: Analysis Without Time Shows a Significant Treatment Effect.



6.3.2.4.1 Simulation of Treatment

Overall detection rates of a significant treatment effect were 75-80% in each analysis in line with sufficiently powered experiments. In no case was a significant treatment effect favoring the control group detected. For a total of 269 simulations the analyses disagreed on the significance of the block*group interaction. As we expected an improvement, we have looked at one case each, where at least one analysis did not detect a treatment effect. These were selected by the maximal individual p-values and the maximal sum of p-values (figure 19). Similar to the examples presented before, these cases suggest that an improvement is not always observable, but also that an analysis without time can lead to a wrong interpretation of treatment effects. Effects might wrongly be attributed to a treatment if improvements within the pre- or posttest are large in the treatment group. On the other hand, they might wrongly be disregarded if improvements within the pre- or posttest are large in the control group.

6.3.2.5 Simulation of no Treatment

For the simulations without a treatment, all analyses show comparable type 1 error rates (.05-.06)[6] and differences in the detection of effects for the simulation of a treatment are thus not only due to varying type 1 error rates. For almost all detections of treatment effects (105 out of 108), the analyses disagreed on the significance of the block*group interaction. As before, we have looked at one case for both directions of the disagreement (selected by minimal p-values, figure 20). While any detection of a treatment is a type 1 error, these cases suggest a wrong detection of treatment effects due to practice effects for the traditional analysis. With the incorporation of time, treatment effects are detected because of actual fluctuations of performance between blocks.

6.3.3 Other Applications

First, the approach is easily applicable to analyses of improvements in relative performance, improvements with the number of tasks finished instead of time spent on tasks, and, by the use of generalized mixed models, also for categorical response data. Changing the analysis to only number of tasks might not be advisable at this point as we have seen that slower participants improve equally much or more while handling fewer stimuli.

The demonstrated approach is of course also applicable to nonlinear changes over time within sessions, which can be approximated using polynomials or generalized additive models (Winter & Wieling, 2016). However, these cases might not be the most interesting for investigation of treatments in pre-post designs as treatment effects are less clear in meaning if there are already complex practice effects within sessions. Nevertheless, the demonstrated approach can be applied and moreover be adapted to compare different sessions at different time points. While in the case of linear improvements we set time 0 to the end of the pretest and the beginning of the posttest,

[6] This suggests that the choice of random slopes is overall acceptable. However, we do note that closer inspection of the distribution of p-values indicates a possibly higher type 1 error rates for the time analysis. We also note large fluctuations in power and type 1 error rates for different choices of random slopes. While this does not contradict the points discussed here, it supports the importance of the choice of random slopes.

for quadratic performance changes over time, time 0 might be more appropriately set closer to the peak performance within sessions.

Another application of the demonstrated approach could be piecewise linear approximation of changes over time within sessions. Compared with growth curve analyses and generalized additive models, this would be advantageous if effects are expected to be limited in time. For example, improvements due to task familiarization might only affect the first few minutes of a test and fatigue might only start to decrease performance after some time has passed. Piecewise linear approximation allows these effects to be restricted in time. An additional advantage is that coefficients are easily interpretable. However, there are further advantages and disadvantages of local versus global procedures, similar to considerations of piecewise interpolation or segmented regression compared with their global counterparts.

6.4 Discussion

The use of linear mixed models provides the opportunity to include practice effects within sessions in the analysis. We presented an approach, which through inclusion and manipulation of time within blocks can account for these practice effects instead of misinterpreting them as treatment effects. The usefulness of such an approach is demonstrated through examples. Through further random samples drawn from real data, it is demonstrated that such cases are not unrealistic.

With the analysis of performance within one group, we have shown that time can explain performance differences between repeated tests but also demonstrated the possibility to find treatment effects. By comparison with control groups, treatment effects can also be identified without an analysis of time if all groups show comparable performance. If groups differ in performance or, in the using traditional methods undetectable worst case, in learning speed, analyses without time might fail or wrongly attribute treatment effects. In these cases, the inclusion of time in the analysis provides significant improvements for the detection and interpretation of effects. While the examples were constructed from extreme values, their average performances also include cases which are conventionally interpreted as signs of (un-)successful treatments, such as differing or comparable pretest performance and comparable posttest performance between groups. At least in the examples, these conclusions would be wrong.

Overall, the inclusion of time in the analysis has shown to improve the analysis and detection of learning and treatment effects. The improved analysis of course comes with the cost of higher complexity. Using random simulations of group allocations, we have shown that differences in learning speed within sessions may pose a serious concern and influence regarding interpretation is not limited to constructed cases as over one quarter of the simulations was affected. This suggests that the increase in complexity is rewarding. We do note that in our sessions the treatment and testing sessions were comparable and the practice and treatment effects should be similar. In realistic experiments, the variance and magnitude of practice effects could be smaller compared with treatment effects. However, this is rarely investigated.

While the suggested approach approximates practice effects, this is limited to effects within sessions and thus does not lessen the need for control groups, which can also account for other improvements between sessions. However, there is no consensus on which control activities optimally control for changes between sessions (Au et al., 2020). As such, the proposed analysis can supplement the control of placebo effects, especially if the comparability between control and treatment groups is hindered by performance or learning differences between groups. An advantage compared with the use of control groups is that no additional lab time or participants are required. This could be applicable, for example, in pilot studies or investigations of details of known treatments, where resources could better be invested in increasing power.

6.5 Limitations

While it allowed the construction and demonstration of multiple different cases, one major limitation is the simulation of groups and blocks from a dataset that did not originally include these. As all datasets were created from data of a single testing session there were no breaks between blocks, which might aid the linear continuation over time. This might transfer well to short experimental sessions where pre- and posttests are conducted within one hour. For multiple sessions spread out over days or years, the results might not be as applicable. But in these cases, the analysis of learning effects during single sessions and the approximation of performance toward the end of sessions might be of interest.

The demonstrated approach is also less useful for tasks, which produce only small practice effects within tests but larger effects between tests. Compared with quick adaptations in cognitive or coordinative tasks, for example, strenuous physical exercise relies on rather slow morphological adaptation. While the investigation of changes within sessions might nevertheless be interesting, the incorporation into statistical analysis can only marginally control practice effects in such a case.

6.6 Conclusion

The use of linear mixed models provides the opportunity to include practice effects within sessions in the analysis. Through constructed examples and random simulations drawn from real data, we have shown the beneficial effects of the inclusion and manipulation of time in the statistical analysis with linear mixed models. The statistical model is improved, allows a more accurate analysis of learning effects, and a better detection of treatment effects. The proposed analysis can supplement comparisons with control groups and aid interpretation of treatment effects especially if such comparisons are hindered by pretest differences between groups.

7 Fourth Study: Manual and Visual Training of Mental Rotation[7]

In this study, we turn back to motor processes as in the first study. Here, we are concerned with the specific similarity of mental rotation and manual rotation rather than general exercise. As the interpretation of mental rotation reaction time is linked to a rotation speed as in physical rotation, exploring manual rotations can facilitate our understanding of the mental rotation processes. Manual rotation interventions have shown training effect for mental rotation performance, but the underlying mechanisms are not fully understood. Here, we apply the design of the second study and the analysis of the third study to analyse three training interventions. Moreover, we can further evaluate the design of the second study.

7.1 Goal and Hypotheses

Due to the lack of separation of the visual and motor component in manual rotation trainings, it is the main goal to determine which component of the manual rotation training influences mental rotation performance. Based on the mental rotation experiment of the second study, we conducted mental rotation tests and manual trainings, in which both the visual component and the congruency of the motor component were separated. Compared with other manual rotation trainings employing a two-figure design, which can only use congruent stimuli, the design of the second study can be used for both mental and manual rotation. By using the same design for the training, more direct comparisons between training and tests are possible. However, the design itself needs to be evaluated further, as it has not been widely used yet.

The following three training conditions were investigated: the "wheel" training comprises the manual rotation of stimuli using a steering wheel (causal and congruent motor activity for visual rotation), the "button" training the manual rotation of stimuli using button presses (causal but not

[7] The results presented in this chapter were published in advance in: Jost, L., & Jansen, P. (2021). Manual training of mental rotation performance: Visual representation of rotating figures is the main driver for improvements. Quarterly Journal of Experimental Psychology. https://doi.org/10.1177/17470218211039494

congruent motor activity for visual rotation), and the "visual" training the automatic visual rotation of stimuli (no causal motor activity). The following primary hypotheses are:

H1) All training conditions improve mental rotation performance.

H2) The "wheel" training shows a larger training effect than the "button" training due to the congruent motor activity and both "wheel" and "button" training show a larger training effect than the "visual" training due to the causal motor activity.

As secondary hypotheses, we investigate at first, if the effects of the test design can be replicated and secondly, the relevance of gender and experience. Between participants, the effect size of gender differences varies between null or small effects (*d* = 0-0.45) in chronometric mental rotation tests and large effects (*d* > 0.7) in psychometric mental rotation tests with better performance of male participants compared with female participants (Jansen-Osmann & Heil, 2007b; Voyer et al., 1995; Voyer & Jansen, 2016). In the second study, gender differences were not studied. As it resembles both classical chronometric and psychometric tests, we expect at most medium effect sizes (*d* = 0.5) and we expect these to diminish in the posttest if pretest differences occur. Effects of training and previous experience are large (*d* > 0.7) and persist throughout multiple sessions (Jost & Jansen, 2020; Meneghetti et al., 2017). Thus, we expect participants with previous experience with mental rotation to perform better in both pre and posttest.

The following secondary hypotheses are investigated:

S3) Regarding the test design: We expect an improvement of mental rotation performance over time with larger improvements for larger angular disparities. Furthermore, performance differences by the side of the correct answer and by the axis of rotation are expected.

S4) Regarding differences by gender and experience: Male participants will perform better than female participants. Participants with experience will perform better than participants without experience. We expect the worse performing group to improve more between pre- and posttest.

Moreover, we preplanned exploratory analyses on differences in training effects between the trained axes and the stimuli used during the training session and the pretest as well as the effects of training parameters (such as planning times and rotation speed) on posttest performance. Previous research on both mental and manual training found both improvement only on the trained axes and models, that is, instance-based learning (Heil et al., 1998; Wiedenbauer et al., 2007) as well as transfer to unlearned models and axes, that is, process-based learning (Adams et al., 2014; Wiedenbauer & Jansen-Osmann, 2008; Wright et al., 2008). Because the manual rotation allows further characterization by training parameters and Adams et al. (2014) found that these parameters improved differently after training, we want to explore if and how these influence mental rotation performance. This could also help to identify the aspects of the training which are most important for its effectiveness.

7.2 Method

7.2.1 Participants

While overall effects of a manual training of mental rotation are typically moderate to large compared with a control group (Adams et al., 2014: η_p^2 = .13-.18; Wiedenbauer et al., 2007: *d* = 0.66) effect sizes between different types of manual training are unknown but expected to be lower. Because repeated practice of mental rotation has shown improvements over larger training volumes (Meneghetti et al., 2017), we assumed larger effects for longer training sessions and employed a training of about twice the volume compared to Adams et al. (2014) and Wiedenbauer et al. (2007)[8]. We estimated the total required number of participants at 192 (64 per training condition) in G*power (Faul et al., 2007). This should yield appropriate power of .8 for medium effect sizes of *d* = 0.5 at the standard .05 alpha error probability for all pairwise comparisons between groups and for small effect sizes of *f* = 0.11 for the within-between interaction of groups and their

[8] We estimated about 360 trials per participant in the 30-minute training session based on an average reaction time of 4 seconds, and 1 second break between stimuli compared with 176 and 192 training stimuli.

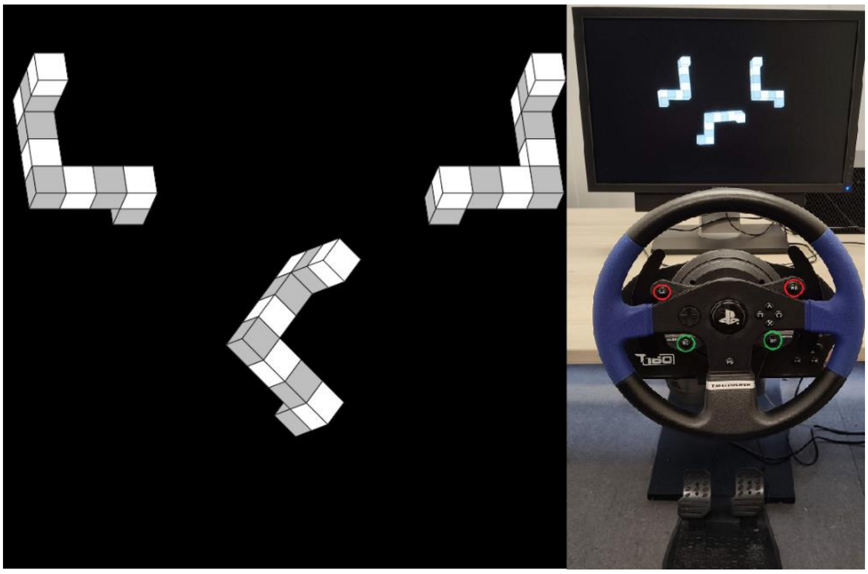
improvements. This should also suffice for appropriate power regarding the within-subject effects of mental rotation for the secondary hypothesis S3. Brysbaert and Stevens (2018) suggest at least 40 participants with at least 40 trials per condition which we should exceed with our design. For the secondary hypothesis S4 regarding gender and previous experience estimated effect sizes are small to medium (*d* = 0.4) for gender and large (*d* = 0.7) for previous experience. While we did not target participants of specific gender or experience level, we expected a somewhat symmetric distribution. Assuming at least 64 participants in the smaller of the two groups in each case, G*power (Faul et al., 2007) shows sufficient power for the analysis.

Participants were recruited by advertisement in the newsletter for students (Bachelor Applied Movement Science) at University of Regensburg and received study credit for participation. They were required to physically be able to press pedals with their feet and use their hands to turn a steering wheel. Other than that, there were no exclusion/inclusion criteria.

Due to the outbreak of the COVID-19 pandemic, testing was interrupted before the targeted number of participants was reached. The global pandemic was likely to cause disruptions in education, physical, and psychological well-being as well as other effects during the large breaks in time of testing (such as an increased scientific interest in medicine or increased use of digital devices), which we expected to affect mental rotation performance and increase variance within groups but did not expect to alter the relative effectiveness of trainings. By including participants both before and after the pandemic, additional confounding variables would have to be analyzed. If such a new analysis were necessary, it seems beneficial to also incorporate findings of present results to improve the experimental design and target more specific research questions. Thus, we decided to analyze and publish the results of the already tested 121 participants ("wheel" training group: *N* = 38, 12 men and 26 women, "button" training group: *N* = 42, 17 men and 25 women, and "visual" training group: *N* = 41, 11 men and 30 women). The mean age was 21.4 years, (*SD* = 1.9) and did not differ significantly between groups. Data from further three participants was incompletely recorded due to electrical failures and other programs interfering during the experiment and not analyzed. The desired power for the main hypothesis is thus only achieved for effect sizes of d = 0.62-0.63 for pairwise comparisons and f = 0.14 for the within-between interaction. Nevertheless, we deemed the analysis interesting both for insight into the achieved training effects as well as modifying the experiment for future research. The use of mixed models for statistical analysis should also enhance power compared to the power analysis for t-tests and ANOVAs. Furthermore, the additional Bayesian analysis indicates sufficient evidence for the main hypotheses.

Figure 21

Mental Rotation Task and Manual Rotation Setup.



Note. Left: Mental rotation task. The top two figures are the base figures, which are mirrored to each other. The stimulus in the center is a rotated version of one of the two base figures and participants are tasked to identify the congruent base figure. Right: Experimental setup including pedals and steering wheel, which participants could use to rotate the stimulus in the "wheel" and "button" condition. The buttons used are circled red (L2 and R2, upper two buttons) and green (SE and ST, lower two buttons).

7.2.2 Material

7.2.2.1 Mental Rotation

Stimulus presentation and response handling were controlled with OpenSesame software (Mathôt et al., 2012) on a Dell OptiPlex 7050 Tower stationary desktop with a Dell P2210 screen (22", 1600x1050, 60 Hz). The screen was placed approximately 40cm from the edge of a desk with a Thrustmaster T150RS steering wheel attached centrally in front of the screen and the according two foot-pedals placed approximately 40cm from the edge of the desk on the floor, both using the more resistant spring of the brake pedal. The pedals were placed against a metal plate to prevent movement backwards (see Fig. 21). The internal forces of the steering wheel were scaled to 50% and set to return to the neutral position. This allowed turning the wheel with little force and a reliable return to the neutral position when the wheel was not held. Participants were seated in a wheeled office chair and were free to adjust their seating position.

Cube figures were used as stimuli and were generated from the stimulus library of the second study with the parameters given in table 18. The mental rotation task was presented according to the layout of the second study with two base figures at the top and one stimulus figure below with a vertical shift of 150pixels (above the center of the screen for the base figures and below the center for the stimulus figure) and a horizontal shift of 300 pixels (left and right of the center of the screen for the base figures). The resulting task is shown in Figure 1.

Trials were shown until a response was given and after every trial the participants received feedback for 1000ms (✓- right, ✗- wrong) shown at the center of the screen at font size 40. The next trial did not start if a pedal was pressed, or the wheel was turned by more than 25° during the "wheel" training condition. Then, a "+" was shown at the center of the screen until all pedals were released. The order of stimuli was block randomized at the start of every part of the experiment using 20 blocks, such that in each block, each unique combination of eligible stimulus properties occurs only once. No stimulus occurs both in the last 10 stimuli of one block and in the first 10 stimuli of the following block. Each part of the experiment was limited by time such that the maximal number of stimuli was never reached.

Participants were instructed prior to the first trial by on-screen text to press the left foot pedal with their left foot if the stimulus could be rotated into congruence with the left base figure. If the stimulus could be rotated into congruence with the right base figure, participants should press the right foot pedal with their right foot. Pedals had to be pressed a minimum of halfway down to register. Participants were asked to answer as quickly and as precisely as possible for both the pre- and the posttest but not for the training session.

Table 18

Parameters for Generation of Stimuli.

Parameter group	Parameter	value
Color options	Background color	Transparent (black)
	Border color	black
	Face color	Grey, white
Sizing and formatting	Cube Diameter	50px
	Image size	440px*440px
	File format	png
	Centering	Optical
Model properties	Base orientations	a,b
	Models	Peters and Battista (2008), 1-16
	Base rotation angles (x,y,z)	-15°,0°,15°
	Angle difference	45°
	Rotation axes	y,z

In the pre- and posttest, participants were instructed to rotate the stimulus figure in their mind and select the base figure that is congruent to the stimulus figure.

In the “wheel” training condition, participants were instructed to turn the stimulus figure using the steering wheel into congruence with one base figure. The rotation of the stimulus was updated at every computational step to the position of the steering wheel rounded to the nearest 3°. Participants were instructed to release the wheel between trials to return to the neutral position.

In the “buttons” training condition, participants were instructed to turn the stimulus using the buttons on the steering wheel. They had to simultaneously press the two top buttons (L2 and R2) to turn the stimulus clockwise and the two bottom buttons (SE and ST) to turn the stimulus counterclockwise (Fig. 21). They had to press two buttons to avoid congruence or incongruence between the pressed buttons and the base figures. The stimulus was turned 3° in every computational step as long as the buttons were pressed.

In the "visual" training condition, participants were instructed to watch the stimulus turn into congruence with one base figure. After showing the stimulus for 500ms the stimulus turned 3° in every computational step until congruence was achieved. The direction of rotation was always the shortest path and random for starting angles of 180°. Due to a programming error, the stimulus was already turned 3° before the first showing.

As computational steps took about 60ms on average, the rotational speed was about 50°/s in the button and visual training condition which is comparable to the mental rotational speed of the second study (64°/s) and the manual rotation speed of Wiedenbauer et al. (2007: 43.48°/s).

In all training sessions, answers were only allowed after a discrepancy of at most 9° between the stimulus and the correct base figure was achieved at least once. Training sessions only used rotations around the y-axis (the picture plane) and no starting angles of 0°.

Reaction time, accuracy, stimulus type (model, angular disparity, rotational axis, stimulus orientation, base orientation), and time since start of each part of the experiment were recorded for all trials. During the trials the rotation of the steering wheel, the state of the relevant buttons of the steering wheel, the position of the pedals, and the shown angle of the stimulus were recorded for every computational step.

7.2.2.2 Demographics

A digital questionnaire was used to collect demographic information. Participants were asked about their previous experience with mental rotation (participants had to indicate if they had or had not participated in other mental rotation experiments before, yes/no), age (in years), gender (male, female, or diverse), information about their menstrual cycle, physical and musical activity, and handedness. Besides previous experience and gender, these were not part of the analyses in line with the preregistration.

7.2.2.3 Procedure

Participants completed a pretest of 10 minutes followed by a training session for 30 minutes, a posttest of 10 minutes, and a digital questionnaire. Between all parts, participants had a self-paced break with instructions for the next part. Participants were informed before the start of the experiment about the length of each part and were shown again before each part. For the duration of the experiment and the questionnaire, participants were alone in the experimental room. All parts were controlled by time and not by the number of stimuli to ensure a comparable overall duration of the experiment between participants.

Participants were randomly assigned to the training conditions using block randomization with one block generated by random sampling in R (R Core Team, 2018). For each of them, the base models of the stimuli were randomly selected using shuffling in OpenSesame (Mathôt et al., 2012) such that two unique models were used exclusively in each of the three parts and two unique models were used in each combination of two parts but not in the third. As a result, six different models were used in each part and twelve models were used in total for each participant. The remaining four models were not used. Models were randomly selected for every participant to avoid influences of systematic differences between models.

7.2.2.4 Statistical Analysis

The accuracy and response time of each trial were used as dependent variables and the training condition (group), the angular disparity, time (since start of each part, within each part) and block (pretest or posttest), the side of the correct answer, the rotational axis, and gender and previous experience of participants were used as independent variables. Angular disparity and time were treated as continuous variables and the categorical variables used treatment contrasts. Contrary to our preregistration the angular disparity was calculated for each rotational axis separately. This allowed us to include non-rotated stimuli, for which the rotational axis is not well defined, in the analysis of axes. Given that an improvement over time is expected and that this improvement is expected to be larger for larger degrees of rotation, the four-way interaction degree*time*group*block for each axis was analyzed for the main hypotheses. Here, the effect of time represents the improvement within the pre- and posttest, whereas the effect of block describes the improvement between tests, i.e. the treatment effect. We expected an improvement by block, which exceeds the expected improvement over time for all conditions, and the interaction of block*group to explain differences in improvement between groups. For the secondary hypothesis we analysed the interactions degree*time*side for each axis, gender*block, and experience*block.

Outliers were determined for each rotation angle by a deviance of more than three standard deviations from the mean reaction time of all stimulus pairs with the same rotation angle and were excluded from all analyses. Reaction time was additionally only analyzed on correct responses. By this procedure, 1178 of 85354 trials (1.3%) were deemed as outliers (455 of 16162 in the pretest, 559 of 49119 in the training, 164 of 20073 in the posttest). Of the remaining trials, 6766 responses (8.0%) (2739 in the pretest, 1477 in the training, 2550 in the posttest) were incorrect.

Statistical analysis was performed as in the first studies with linear mixed models using lme4 package (version 1.1-21; Bates, Mächler, et al., 2015) in R (version 3.5.1; R Core Team, 2018). Model parameters were estimated by maximum likelihood estimation using bobyqa algorithm wrapped by optimx package (version 2018-7.10; Nash & Varadhan, 2011) as optimizer. Model fit was calculated by using likelihood ratio tests to compare models with and without the fixed effect of interest. The resulting p-values were compared to a significance level of .05. Visual inspection of residual plots did not reveal deviations from homoscedasticity or normality in any model.

For the significant effects of interest, we report both the unstandardized effect sizes and confidence intervals calculated by using parametric bootstrapping with 1000 simulations in line with recommendations of Baguley (2009) and Pek and Flora (2018). While standardized effect sizes are routinely used for power analysis and meta analyses, unfortunately there does not exist an agreed upon way to compute standardized effect sizes in linear mixed models (Feingold, 2009; Hedges, 2007; Rights & Sterba, 2019). Nevertheless, linear mixed models offer several advantages over traditional use of ANOVAs. For example, linear mixed models allow simultaneous analysis of by-participant and by-item variances and thus eliminating the need to average over participants or items, while also allowing analysis of unbalanced data and achieving higher statistical power (Barr et al., 2013; Hilbert et al., 2019).

Model building was based on the research of Barr et al. (2013), Bates, Kliegl, et al. (2015), and Matuschek et al. (2017), starting with a model with random intercepts and slopes for every appropriate fixed effect and reducing the model complexity by dropping non-significant variance components (to avoid over-parameterization at the start we included all two-way interactions for random slopes by participant and only the main effects for random slopes by the stimulus model). Non-significant fixed effects were further stepwise removed from the model, such that effects which least decreased model fit were removed first and a model containing only significant fixed effects remained. Non-significant effects were then tested for an improvement of model fit by inclusion in the resulting model, while significant effects were tested for worsening of model fit by exclusion of the effect. The resulting models for each parameter are described in the results section. The analysis of numerical main effects contained in significant interactions was performed according to Levy (2014). Degree was centered such that main effects show the average improvement over all angles. Time was normalized such that time 0 was set to the end of the pretest and the start of the posttest. As a result, the effect of block represents the difference between the estimated end of the pretest compared to the beginning of the posttest. While the comparison between the groups was the main goal of the study, this inclusion of time in the analysis allows for better control of practice effects within the tests. Thus, the treatment effect of interest are the block*group interactions.

Due to the non-significance of many results we have retrospectively calculated Bayes factors to distinguish evidence in favor of no effects (Dienes, 2014; Wagenmakers, 2007). Due to the retrospective nature of the analysis we opted to calculate Bayes factors objectively based on the approximation of Wagenmakers (2007) and compared it to the decision boundary factor 3 or $\frac{1}{3}$. We do note a monotonous relationship between the Bayes factors and p-values and thus the necessity to consider the Bayes factors also for significant results, which we elaborate on in the appendix.

For both frequentist and Bayesian analyses there is ongoing discussion about the optimal procedure and we release all data and code in accordance with the suggestion of Matuschek et al. (2017).

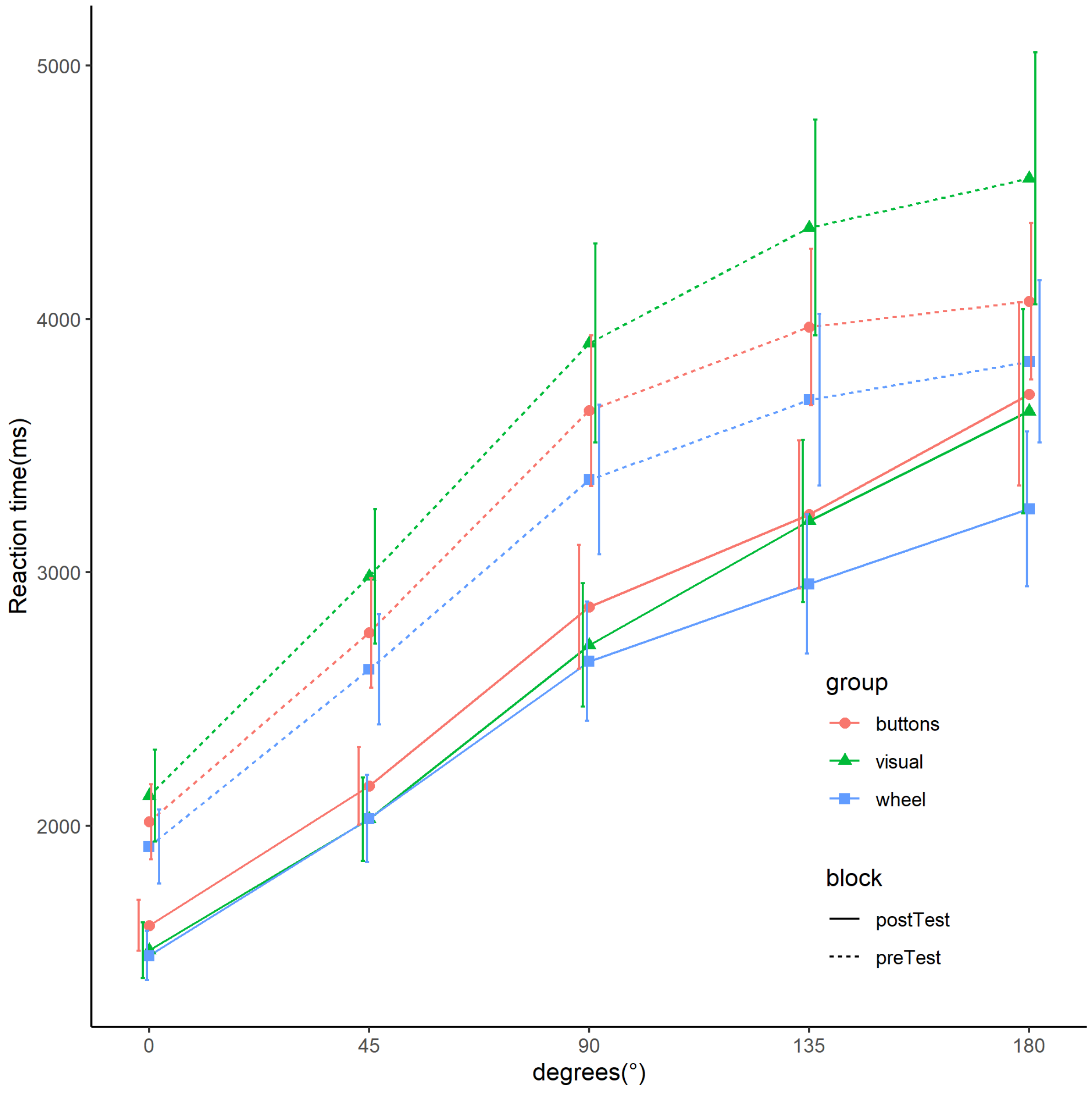
7.3 Results

7.3.1 Descriptive Statistics

Summarized performance data is shown for reaction time (Fig. 22) and accuracy (Fig. 23). Due to the time-controlled nature of the experiment, participants finished different number of mental rotation trials. To account for this, mean data is first calculated for every participant and then averaged over participants. Further summaries of behavioral data and summarized demographic data can be found at https://github.com/LeonardoJost/MMR.

Figure 22

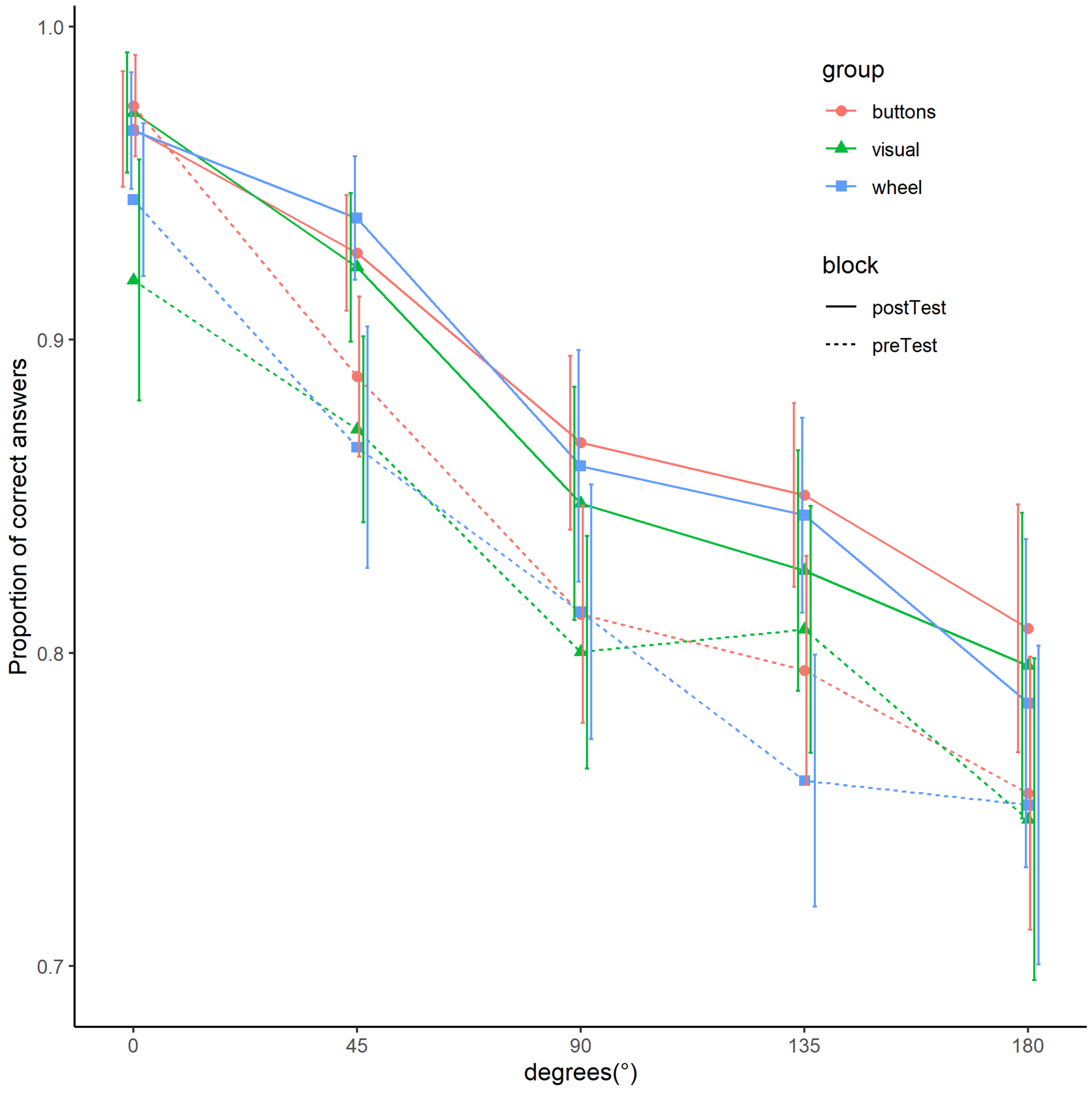
Mean Reaction Time of Mental Rotation Trials as a Function of Angular Disparity for the Three Groups and Two Tests.



Note. Mean reaction time is calculated for all correctly answered trials of every participant and then averaged over all participants. Error bars show 95%CIs computed by ggplot2 (Wickham, 2016).

Figure 23

Mean Accuracy of Mental Rotation Trials as a Function of Angular Disparity for the Three Groups and Two Tests.



Note. Mean accuracy is calculated for every participant and then averaged over all participants. Error bars show 95%CIs computed by ggplot2 (Wickham, 2016).

7.3.2 Comparison of Pre- and Posttest

7.3.2.1 Reaction Time

For the analysis of reaction time, model building resulted in a model with random intercepts and random slopes for degree, time (since start of each part), and block by participant and random intercepts by model. Significant effects were found for degree(y-axis)*time*block*group, degree(z-axis)*block, degree*side, degree*time, gender*block, and experience*block.

Overall and in line with hypothesis H1, there was an improvement for all stimulus properties (degree and axis), which only differed in magnitude. This was supported by both frequentist and Bayesian analysis. The significant four-way interaction was inconclusive regarding the Bayes factors suggesting possible overall differences between groups in learning over time of rotations around the y-axis within the pre- and posttest. However, no partial interactions containing the block*group interaction proved significant contrary to hypothesis H2. The Bayes factors indicated strong evidence for no effects. This suggests a comparable overall treatment effect (the block*group interaction), comparable overall improvements of rotations around the y-axis (the deg(y)*block*group interaction), and comparable changes of learning within tests (the time*block*group interaction) between groups.

In line with our secondary hypothesis S3 improvements over time were larger for larger angles and answers on the right side showed a lower slope by degree compared with answers on the left side but no main effect between the sides proved significant. The Bayes factors were in support of the significant results and suggested null effects for the non-significant results. The frequentist and Bayesian analysis disagreed on the difference in improvements over time between axes, suggesting the necessity for further research. Contrary to the secondary hypothesis S4, neither main effects of gender nor experience were significant, but men improved more than women and participants without experience improved more than participants with previous experience. The Bayesian analysis supported the larger improvement for men and suggested no overall gender differences but required more evidence for both effects of experience.

Regarding the exploratory comparison of training effects between axes, rotations around the z-axis showed a steeper slope by degree at the start of the posttest compared to the end of the pretest, whereas there was no significant difference for the y-axis. This was partially supported by the Bayes factors, requiring more evidence for the z-axis. These results partially suggest a larger training effect for rotations around the y-axis but due to the overall improvement between blocks also a training effect for the z-axis.

For the interactions by block, we conducted separate analyses for the pre- and posttest. Participants with experience were significantly faster in the pretest but not in the posttest. Despite the difference in training effect, gender differences were not significant, in neither pre- nor posttest. The decomposition of the four-way interaction revealed significant axis differences in the pretest and between-group differences regarding the axes and improvements over time in the posttest. Bayes factors were inconclusive regarding the effect of experience and differences in improvement over time in the posttest. For the between-group differences regarding the axes, the Bayesian analysis contradicted the frequentist analysis suggesting evidence of no differences. (see table 19)

Table 19

Statistical Analysis of Reaction Time.

Variable	Estimate	SE	Test statistic	p	95% CI	BF
Intercept	2796.21	118.16			2562.43, 3007.02	
Deg(y)*time*block*group			$\chi^2(2)=8.17$	.017		2.04
Deg(y)*block*group			$\chi^2(2)=0.41$	.816		98.73
Time*block*group			$\chi^2(2)=2.04$	.361		43.64
Block*group			$\chi^2(2)=2.05$	.359		43.40
Block(pre-post)	242.91	52.29	$\chi^2(1)=13.44$	<.001	141.46, 351.74	0.01
Deg*time*side			$\chi^2(1)=3.71$	.054		1.72
Deg(y-z)*time			$\chi^2(1)=3.90$	.048		1.56
Deg(z)*time	-364.65	149.92	$\chi^2(1)=5.91$	.015	-649.91, -63.27	0.57
Deg*side(right-left)	-107.42	26.00	$\chi^2(1)=17.06$	<.001	-157.13, -55.48	<.01
Side(right-left)	10.03	13.41	$\chi^2(1)=0.56$	.455	-16.66, 36.23	8.32
Block(pre-post)*deg(y)			$\chi^2(1)=0.31$	.578		9.42
Block(pre-post)*deg(z)	-133.02	56.60	$\chi^2(1)=5.52$	.019	-243.76, -14.31	0.70
Block(pre-post) *gender(female-male)	-246.34	76.52	$\chi^2(1)=9.70$	.002	-403.35, -103.40	0.09
Block(pre-post) *Experience(no-yes)	338.13	132.44	$\chi^2(1)=6.14$	.014	67.27, 609.69	0.51
Gender(female-male)	28.17	75.81	$\chi^2(1)=0.02$	.875	-107.39, 184.71	10.87
Experience(no-yes)	210.18	131.65	$\chi^2(1)=3.71$	.054	-57.25, 467.33	1.72
Pretest						
Gender(female-male)	-153.83	111.62	$\chi^2(1)=1.81$	.178		4.44
Experience(no-yes)	643.10	199.01	$\chi^2(1)=9.58$	.002	216.62, 1002.14	0.09
Deg(y-z)	61.88	20.99	$\chi^2(1)=8.68$	.003	21.42, 102.42	0.14
Posttest						
Gender(female-male)	-35.01	70.17	$\chi^2(1)=0.25$	.620		9.71
Experience(no-yes)	204.56	120.93	$\chi^2(1)=2.77$	.096		2.75
Time*group(buttons)	-463.20	276.12	$\chi^2(2)=8.79$	.012	-1004.78, 67.88	1.49
Time*group(visual-buttons)	-1176.73	390.80			-1921.63, -425.25	
Time*group(wheel-buttons)	-647.79	397.14			-1394.02, 138.12	
Deg(y-z)*group(buttons)	-216.04	26.03	$\chi^2(2)=6.50$	.039	-268.58, -161.78	4.69

Variable	Estimate	SE	Test statistic	p	95% CI	BF
Deg(y-z)*group(visual-buttons)	77.82	36.93			6.09, 149.31	
Deg(y-z)*group(wheel-buttons)	-8.01	37.00			-81.57, 65.68	
Exploratory						
N(pretest)*group			$\chi^2(2)=4.27$	.118		14.33
N(pretest)	-8.09	0.76	$\chi^2(1)=51.79$	<.001	-9.51,-6.52	<.01
N(training)*group(buttons)	-2.57	1.10	$\chi^2(2)=7.85$	.020	-4.65, -0.39	2.39
N(training)*group(visual-buttons)	-2.96	1.54			-6.19,-0.08	
N(training)*group(wheel-buttons)	0.60	1.19			-1.69, -2.71	
N(training)			$\chi^2(1)=23.57$	<.001		<.01
Proportion(short direction)*group(buttons)	410.19	392.36	$\chi^2(2)=5.00$	.025	-346.79, 1200.04	9.93
Proportion(short direction)*group(wheel-buttons)	-1650.71	726.36			-3010.83, -188.41	
Trained models(new-old)	-59.46	17.39	$\chi^2(1)=11.69$	<.001	-91.02, -24.83	0.03
Trained models(pretest-training)			$\chi^2(1)=0.00$	.986		11.00
Trained models*group			$\chi^2(2)=3.31$	.191		23.12

Note. The values for degree and time (since start of each part) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. BF stands for the approximation of the Bayes factor by Wagenmakers (2007) in favor of the null hypothesis.

7.3.3 Accuracy

Accuracy was analyzed by a general linear-mixed model, which used a binomial distribution. Model building resulted in a model with random intercepts and random slopes for degree, time (since start of each part), and degree*time by participant and random intercepts and random slopes for degree and time by model. Significant effects were found for time*block, degree (y-axis)*time*group, and degree (z-axis)*block.

As in the analysis of reaction time, this suggests differences between groups in learning within the tests but no differences in the treatment effect between tests. Overall, there was an improvement from pre- to posttest but again, the rotation around the z-axis showed a steeper slope by degree. Improvements over time were only significant in the pretest and did not differ significantly by degree. No effects of gender or experience were significant. The Bayes factors were in support of the significant results and suggested null effects for the non-significant results except for the inconclusive interaction of gender and block. (see table 20)

Regarding the hypotheses, the results for accuracy were either in the same direction as the results for reaction time or supported null effects. This suggests that the changes in reaction time are not due to speed-accuracy trade-offs.

Table 20

Statistical Analysis of Accuracy.

Variable	Estimate	SE	Test statistic	p	95% CI	BF
Intercept	2.32	0.14			2.04, 2.61	
Block*group			$\chi^2(2)=5.36$	.069		8.30
Deg*time*side			$\chi^2(2)=1.81$	.404		48.95
Deg*time			$\chi^2(1)=0.21$	.644		9.90
Deg*side			$\chi^2(1)=1.72$	.190		4.65
Time*block(pre-post)	1.76	0.33	$\chi^2(1)=27.85$	<.001	1.15, 2.39	<.01
Block(pre-post)*deg(z)	0.36	0.07	$\chi^2(1)=26.84$	<.001	0.23, 0.49	<.01
Block(pre-post)	-0.25	0.06	$\chi^2(1)=15.37$	<.001	-0.38, -0.13	<.01
Time*deg(y)*group			$\chi^2(2)=12.42$	.002		0.24
Gender*Block			$\chi^2(1)=4.03$	.133		1.47
Gender(female-male)	0.09	0.13	$\chi^2(1)=0.47$	.493	-0.17, 0.34	8.70
Experience*Block			$\chi^2(1)=0.55$	.761		8.36
Experience(no-yes)	0.01	0.22	$\chi^2(1)=0.00$	.962		10.99

Note. The values for degree and time (since start of each part) represent estimated changes corresponding to changes of 100° and 30 minutes of testing time. BF stands for the approximation of the Bayes factor by Wagenmakers (2007) in favor of the null hypothesis.

7.3.4 Exploratory Analysis

We performed explorative analyses on the influence of training performance on posttest performance. To begin, we defined and computed metrics to characterize the training performance for which we provide descriptive statistics overall and for the changes during the training session. Subsequently, we have analyzed the relationship between these metrics and posttest performance to explore possible connections.

7.3.4.1 Description of the Training Session

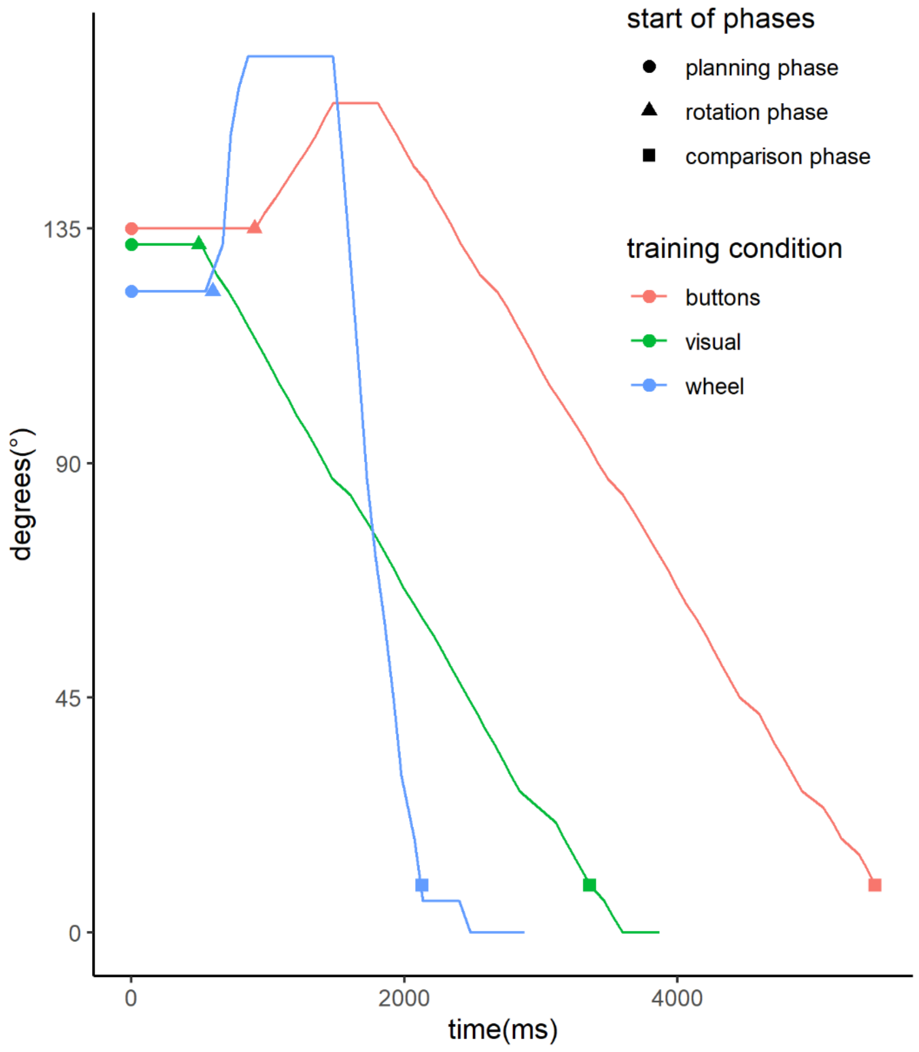
First, the average accuracy for the training trials was above 95% for all three groups and all starting angles, which is comparable to the accuracy of non-rotated trials in the mental rotation test. Next, we have looked at the overall reaction time and other parameters of the training session. Similarly to Adams et al. (2014), we have divided each trial into three phases: A planning phase, a rotation phase, and a comparison phase (see fig. 24 for examples). The comparison phase differs from the fine-tuning phase of Adams et al., as they required participants to match figures as closely as possible whereas our participants were asked to select one of two comparison figures. The planning phase describes the time from the start of the trial to the first angular deviation from the starting position. To account for random fluctuations possibly from returning the wheel to neutral position after the previous trial, the most common angle of the first five measurements (about 240ms) was used and deviations were measured afterwards. The following rotation phase ends when the rotated figure is closer than 10° to the congruent base figure as answers by participants were allowed after this time. The comparison time is the further time until an answer was given. The rotation phase is further described by three parameters: the average rotation speed, the number of switches of the direction of rotation, and whether the overall rotation was performed in the shorter direction (for starting angles differing from 180°). For all parameters, we have descriptively looked at group differences over the course of the training session (Fig. 5). There are differences between groups regarding the phases, both in average values as well as in changes over the course of the session. Notably, overall reaction time and most phases (except for the rotation phase of the "buttons" training and the comparison phase of the "wheel" training) show a steeper decline in the first five minutes compared with the following time in all groups. Regarding rotations in the short direction, average values were 85% for the "buttons" group and 89% for the "wheel" group. This suggests that the planning and rotation phases were used to determine the shortest path of rotation and mental comparison processes are performed before the comparison phase.

7.3.4.2 Analysis of Training Effects

To analyze the training effects of the training parameters, we compared all training parameters averaged by participant and their interaction with group on posttest performance of reaction time. As the training parameters are linked to average performance, we further compared the significant training parameters with the overall number of trials in both the pretest and the training session. Both the number of pretest trials and the number of training trials as well as the interaction of the number of training trials with group were significant. With increasing number of pretest trials and increasing number of training trials the reaction time in the posttest decreased. For the number of training trials, this effect was larger in the visual training group compared with the other groups. Regarding the training parameters, the proportion of rotations in the short direction, and the interaction of the proportion and group showed significant effects on posttest performance. The proportion of rotations in the short direction showed a positive relationship with reaction time in the buttons training group and a negative relationship in the wheel training group. The Bayes factors were in support of the effects regarding the number of pretest trials but inconclusive regarding the number of training trials and contradicting regarding the proportion of rotations in the short direction.

Figure 24

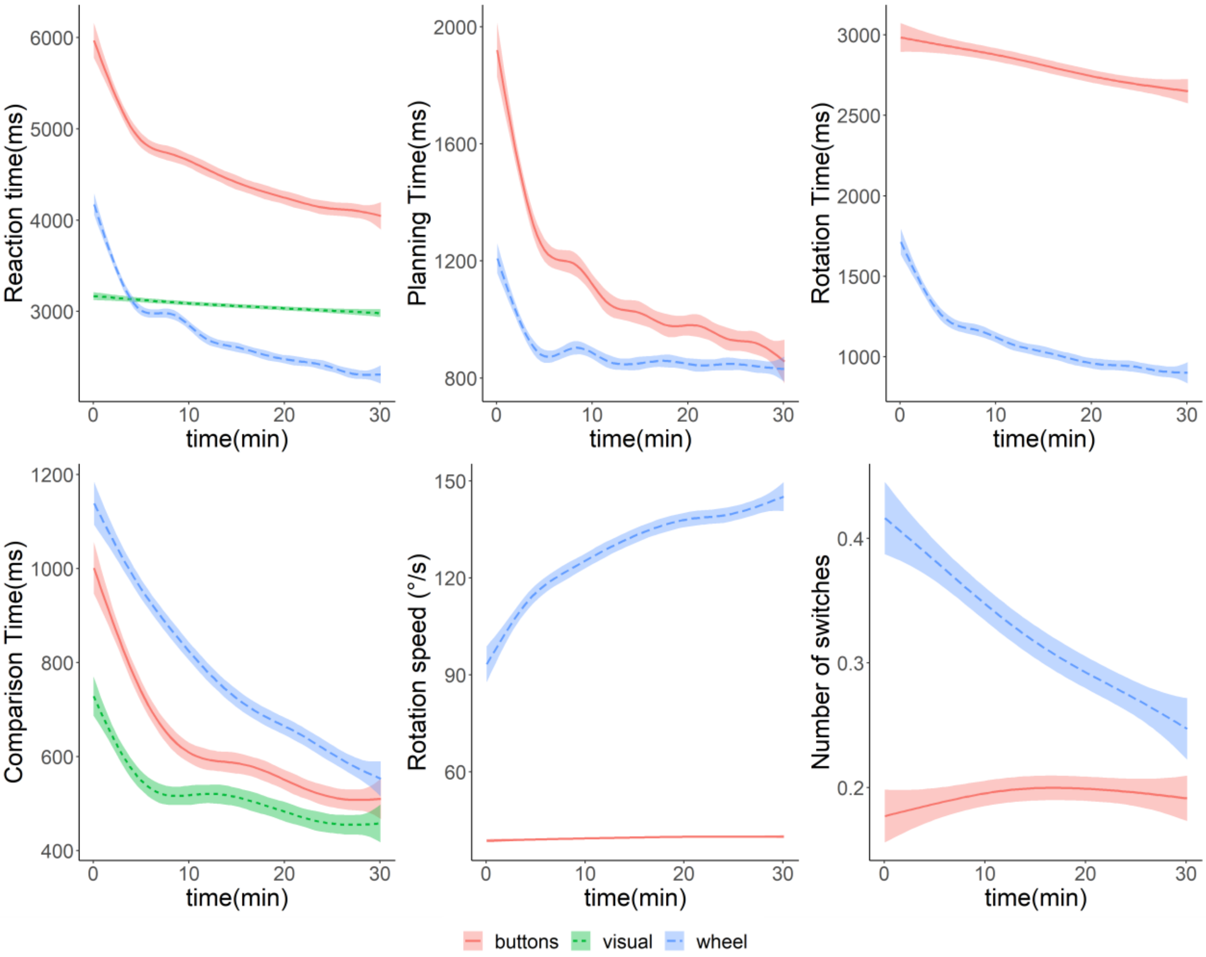
Exemplary Movement of an Object During Training Trials.



Note. The graph depicts one example of a training trial for each condition for a starting angle of 135°. The calculated starts of each phase are marked for each trial. The planning phase starts with the trial onset. The rotation phase starts just before the first angular displacement. The comparison phase starts once a deviation of at most 9° from the target is reached for the first time. The trial ends once an answer is recorded. There are some small fluctuations in the calculated times and the slopes in the "visual" and "buttons" training due to fluctuations in the display frame rate. Because the wheel was not always perfectly aligned and the visual rotation started already turned by 3°, there are small differences in the starting angle.

Figure 25

Parameters of the Training Session Separated by Groups and Their Changes Over Time.



Note. From left to right and top to bottom: (1) reaction time, (2) planning time, (3) rotation time, (4) comparison time, (5) rotation speed, and (6) number of switches of rotation direction. Smoothed conditional means over time are generated using a generalized additive model in ggplot2 (Wickham, 2016).

In a second analysis, we compared the posttest performance of models used in the pretest and in the training session with models only used in the posttest. There was no significant difference between models used in the pretest and in those used in the training session, but both showed significantly faster reaction times than the new models in the posttest. As the differences between models were smaller than differences between blocks, this suggests a transfer of training effect to the new models. Differences between groups were not significant. The Bayes factors were in support of the significant results and suggested null effects for the non-significant results. (See table 19)

7.4 Discussion

In this study, we provide insight into the training effects of manual rotation movements on mental rotation tasks. In line with previous research on manual training of mental rotation (Adams et al., 2014; Wiedenbauer et al., 2007; Wiedenbauer & Jansen-Osmann, 2008) our results show an improvement in mental rotation performance from manual training but these improvements do not differ between a rotational and a non-rotational movement. Moreover, by isolating the component of concurrent visual rotation of stimuli our experiment provides evidence that it is not the motor activity but the concurrent visual rotation that leads to improvements in mental rotation tasks. As repeated mental rotation tasks, where a visual rotation of stimuli in the mind is assumed, show comparable improvements to manual rotation tasks (Adams et al., 2014), this implies that the visual rotation whether internal (imagined) or external (physical or visualized) is the main reason for improvements in mental rotation tasks. The importance of the sensory output is in line with the study of Janczyk et al. (2012). Despite the presentation of visual rotations, the mental representation of this process is not necessarily visual similar to the identified nonvisual representation of mental rotation (Ilan & Miller, 1994; Jansen-Osmann & Heil, 2007a; Liesefeld & Zimmer, 2013).

7.4.1 Training Effects in Mental Rotation

While overall differences between training groups were not significant, a more detailed analysis revealed possible differences in the improvements within the posttest of the slope of reaction time by angle. These might be caused by different parts of the mental rotation process being trained by the training conditions and the repeated mental rotation tasks in the posttest.

Improvements in manual rotation tasks on the other hand might also be driven by familiarization with the motor behavior, which is not trained by the mental rotation tasks. This can be seen by the steep decline in reaction time in the first five minutes of the training session in both manual training groups despite the preceding mental rotation tasks. Improvements afterwards are more in line with improvements found by repeated mental rotation tasks (as in the second study). This supports the results of Adams et al. (2014) that manual rotation performance is improved more by practicing manual rotation tasks than by practicing mental rotation tasks and might also pose a solution for the conflict between their results and the common process hypothesis of Wohlschläger and Wohlschläger (1998).

Regarding the transfer of training effects on new figures, our results show better performance on previously trained figures and a comparable transfer in all groups. This result could be interpreted in support of instance-based learning but the differences between new and old models were much smaller than differences between blocks suggesting also process based learning. The non-significant point estimates of the group differences might indicate a better transfer to new figures with increased motor activity in support of the findings of Adams et al. (2014) but this was not supported by Bayes factors, which showed strong evidence for no effects. Usage of larger differences between stimuli to facilitate these effects might be necessary to generate different transfer effects between objects due to motor activity. Regarding the transfer to untrained rotation axes, our results also show significant improvements for the untrained rotation in depth around the z-axis in all groups. In contrast to the trained axis, this improvement is characterized by a larger improvement on smaller angles. This suggests the transfer of only non-rotational parts of the mental rotation process to the untrained axis.

7.4.2 Performance During Training and Influence on Posttest Performance

Our exploratory analysis of the performance in the training sessions shows that most performance parameters improve during the training and the decreasing reaction times during the training cannot be explained by a single parameter. The performance in the posttest, however, was only significantly influenced by the number of trials, both in the training session and in the pretest. The time spent on trials in the visual training group is mostly programmatically controlled and participants could only influence this time by their response time after stimuli were rotated into congruence. Thus, one could expect that the number of training trials would not be correlated with mental rotation performance in the visual training group or be less correlated than in the other groups. This was not supported in the results, indicating that the time spent on comparing and selecting identical figures is a significant part of mental rotation performance. By another account, the reaction time of the visual training trials might also be influenced by participants rotating stimuli faster mentally than the visual presentation and comparing the stimuli before congruence is achieved. Similarly to the large proportion of rotation in the shorter direction in the manual training groups, this could indicate that mental rotation and comparison processes are performed throughout the training tasks. As Adams et al. (2014) found similar training effects for mental and manual rotation interventions, the facilitation of mental processes by the training could be the most relevant for improvements.

As the reaction time is mostly independent of performance in the visual training, this type of training could be expected to be more suited to slow performers whereas the congruent manual training might be more suited to fast performers, but this hypothesis was also not supported by the results. Pretest performance was a significant predictor of posttest performance, but the Bayes factors indicated no differences between groups.

7.4.3 Implications for Mental Rotation Training

The results suggest the effectivity of a purely visual training to enhance mental rotation performance. This type of training is easy to implement and can easily be adapted for large groups or online training without the need for special equipment or motoric requirements. Such a training can be employed to boost mental rotation ability and spatial ability in general, if such a transfer were found.

Regarding the choice of how to conduct training sessions, the non-congruent manual rotation and the visual rotation would allow further parameterization. If the largest similarity with the congruent manual rotation is desirable, the choice of rotation speed and the starting time for the visual rotation group in our experiment are too low. Furthermore, we observed differences between the congruent and non-congruent manual rotation groups regarding the starting time and the number of switches in rotation direction, which might be caused by the accessibility of the buttons and the simplicity of switching directions. As the improvements were comparable between groups despite these differences, further research is necessary to understand the relationship between the parameterization of the training sessions and their training effects. This also offers the possibility to further optimize and individualize the training.

Compared with repeated mental rotation training, an advantage could be the high accuracy even for complex stimuli. As children have been shown to profit from mental rotation training starting from a young age (Fernández-Méndez et al., 2018) but suffer from the complexity of stimuli (Hoyek et al., 2012), a visual training could help accustom them to more complex tests. However, as Adams et al. (2014) found a similar effect of mental and manual training, the comparison of visual training and repeated mental rotation tasks should be investigated further.

7.4.4 Evaluation of the Mental Rotation Test

For the analysis of the mental rotation design proposed in the second study, our results confirm the proposed similarity to the original chronometric design of R.N. Shepard and Metzler (1971) but also the small left-right differences found in the original study. Reaction time increases and accuracy decreases with degree for both sides and axes and improvements are larger for larger degrees. In support of the need for further research, we also confirmed systematic differences between axes even in the pretest and small differences in the slope between answers on the left and right side. While we did find a significantly larger training effect for men, the non-significance of gender effects in both pre- and posttest is in line with small or non-existent gender differences in chronometric mental rotation tasks (Jansen-Osmann & Heil, 2007b). This was also supported by the Bayes factors suggesting no effects. As expected, the broad measure of previous participation in mental rotation experiments was an indicator of improved performance. The non-significance in the posttest can be explained by the unbalanced distribution of experience compared with our hypothesis. In line with this, the Bayes factors required more evidence for the training effects and posttest performance differences regarding experience. For both gender and experience, the worse performing group in the pretest improved more and reduced the difference to non-significance.

7.5 Limitations

The study is limited by the fact, that there was a different number of trials in the three training groups, though overall time spent was controlled. While it is not clear if time or number of trials is more influential for training effects, this could limit the comparability between groups, which handled different number of trials, and to other studies, which use a fixed number of trials. Due the limited total time and breaks between trials, participants who solved more trials actually spent less time with the stimuli themselves. If the time spent on tasks were the main driver of training effects, our training could have benefitted slower participants more.

Another possible limitation is the fact that the visual training condition is passive, whereas the other two are active. This seems necessary for the separation of the manual and visual components but could interact with training effects.

Furthermore, we could not test 192 participants due to the pandemic interruption, as calculated from the A-priori G-Power analysis. At least for the main hypotheses this should not be a concern as indicated by the Bayesian analysis. The possible training effects regarding previous experience should be treated with caution due to the skewed distribution of participants.

7.6 Conclusion

This study clearly provides evidence that the visual rotation whether internal or external is the most important component for improvements in mental rotation tasks. To isolate the visual component, further investigation of the opposite direction is also necessary. The first step towards this could be the removal of visual rotation in manual rotation tasks, which should result in no or minor improvement of mental rotation performance. While our results support previous findings of manual training of mental rotation (Adams et al., 2014; Wiedenbauer et al., 2007), the design used here is more comparable to the mental rotation task as it employs a congruency judgement similar to the task used by Wohlschläger and Wohlschläger (1998).

Moreover, the study has practical implications for the enhancement or the prevention of a decline in spatial abilities associated with old age (Jansen & Heil, 2009; Meneghetti et al., 2018): A visual training of mental rotation performance might be just as effective as a mental or motor training but the transfer to older adults and other spatial abilities has to be investigated further. Similarly, such a training could be employed for immobile persons such as children with spina bifida who already suffer from reduced mental rotation performance and where a manual training has proven to be effective (Wiedenbauer & Jansen-Osmann, 2007).

Despite the evidence regarding the importance of the visual component in mental rotation tasks and the similarity between mental and manual rotation, the "common process" is far from being understood from an experimental point of view.

8 Fifth Study: Implicit Affective Evaluations and Mental Rotation[9]

As also evaluated in the previous study, sex/gender differences in mental rotation performance are an interesting topic, which we wanted to investigate further. Gender differences in mental rotation are of practical relevance but are still not really understood. One suspected influence is the stimulus material. Recent research at the time investigated the effect of more female stereotyped figures (Rahe et al., 2020; Rahe & Quaiser-Pohl, 2020). Whereas effects of the stimulus material on performance and gender differences have been observed, the underlying mechanisms are not clear. Here, we investigate the rather new and exciting topic of implicit attitudes as a possible explanation as well as some factors already known to influence gender differences.

8.1 Goal and Hypotheses

The main goal of this study is to further investigate stereotypes and social explanations of the male advantage in psychometric mental rotation test performance. Two possible social explanations are the use of spatial toys in childhood (Voyer et al., 2000) and the gender belief that men perform better in mental rotation tasks (Heil et al., 2012), but also the gender stereotype of the stimulus material (Ruthsatz et al., 2014, 2017). Furthermore, it has been shown that implicit gender stereotypes (Guizzo et al., 2019; Hausmann, 2014; Hausmann et al., 2009) and unconscious emotions (Mammarella, 2011) can influence test performance. However, the implicit evaluation of stimulus material in mental rotation tests has not yet been investigated and is a possible source of implicit gender stereotypes and therefore related to mental rotation performance. As we expect the affective evaluation of cube figures to be male stereotyped because of the male attribution of angular designs (Palumbo et al., 2015), we assume that at least the implicit affective evaluation of cube figures is positively related to mental rotation performance. This study will provide evidence

[9] The results presented in this chapter were published in advance in: Jost, L., & Jansen, P. (2021). Are implicit affective evaluations related to mental rotation performance? Consciousness and Cognition, 94(2021), 103178. https://doi.org/10.1016/j.concog.2021.103178

how much the affective evaluations of the objects in a cognitive task are related to the performance in this task.

The following hypotheses will be investigated:

1. First, we assume an implicit (Larson et al., 2012; Palumbo et al., 2015) and explicit negative affective evaluation bias against cube figures but not against pellet figures for women compared to men, measured with an affective priming paradigm. However, if at all, only a small correlation is expected between implicit and explicit measurements (Cameron et al., 2012).

2. Second, we assume to replicate large gender differences in mental rotation performance favoring men compared to women (Voyer et al., 1995) with a larger difference for cube figures compared to pellet figures (Rahe et al., 2020; Rahe & Quaiser-Pohl, 2020).

3. Moreover, the relation between spatial toys, stereotyping of spatial abilities, rating of own spatial abilities, affective evaluation, gender, and mental rotation performance is investigated as well as the prediction of mental rotation performance due to the assumed predictors.

8.2 Method

8.2.1 Participants

With a medium effect size $f = .25$, an alpha-level of $p = .05$, and a power of 1-ß = .95, a power analysis with G*power (Faul et al., 2007) for the repeated measures ANOVA resulted in $N = 54$ (27 men and 27 women) to detect significant interactions of gender and figure type in the explicit as well as the implicit attitudes towards the figures and in the mental rotation performance. With a medium effect size $f^2 = .15$, an alpha-level of $p = .05$, a power of 1-ß = .80, and seven possible predictors for the mental rotation performance of cube and pellet figures (gender, use of spatial toys, stereotyping, self-rating, explicit and implicit evaluations of cube respectively pellet figures, stereotyping of figures), a power analysis for the linear regression resulted in $N = 103$. The number of participants was checked daily and data collection was stopped after 123 participants

(41 men, 82 women) attempted all tasks. Twenty participants were recruited more because possible data must be excluded for performance reasons. We originally intended to exclude all participants, who did not finish all parts of the experiment. However, as missing data could also occur within the experiment, we only excluded the nine participants, who only finished the questionnaires, resulting in 114 participants (39 men, mean age: 22.05 years, *SD* = 2.08, 75 women, mean age: 21.53 years, *SD* = 1.49). Participants were recruited by advertisement in the newsletter for students (Bachelor Applied Movement Science) at University of Regensburg and received study credit for participation. 86% had already participated on some other mental rotation test with another research question.

8.2.2 Material

8.2.2.1 Demographic Questionnaire

Questions concerning age, gender, and participation in former mental rotation tests were asked as well as other demographic data (mother tongue, physical and musical activity), which were not part of the primary analysis.

8.2.2.2 Use of Spatial and Non-Spatial Toys in Childhood

Following the study of Moè et al. (2018), the spatial and non-spatial toys used in childhood were presented to the participants: action figures, blocks, cars and trucks, Lego, video gaming, model kits, puzzle (spatial), and baby dolls, Barbie dolls, dish sets, doll furniture, coloring, board games, and puppets (non-spatial). In comparison to the study of Moè et al. (2018), the spatial game of "ring toss" was exchanged with "video gaming". Participants should indicate on a 5-point Likert-scale how often they have used the respective toy in childhood from 1 (almost never) to 5 (very often). If participants did not know or did not play with one of the toys, they should skip the question, which was given a score of 0. The mean score for the use of spatial toys was calculated.

8.2.2.3 Stereotype-Test

The gender stereotype questionnaire was adopted from Hausmann et al. (2009). Participants were told that they should imagine that they meet a person, who they had never met before and who was either male or female. Participants had to estimate the probability that the individual was male or female for eight sentences of cognitive ability items related to spatial abilities (recognition of complicated drawing, imagination of objects from different perspectives, forgetting the place of common objects, drawing a map, reading a map with street names, using landmarks for orientation, imagine and rotate common objects, remember the way due to right and left turns) and eight items unrelated to spatial abilities. Two columns were aligned next to each item – labeled male and female – and participants entered a number that corresponded to their probability estimate in percentage. If the sum of the probabilities was between 90% and 110% it was normalized to 100% and otherwise discarded from analysis. Because Cronbach's Alpha for the probabilities of the eight items related to spatial abilities was low with .5, we decided in line with Hausmann et al. (2009) but in contrast to our preregistration to include the results of the single questions in the further analyses. Furthermore, differing from the preregistered score, the bias towards the own gender identity was calculated as the difference between the probability of each participant and the overall mean probability and used as a predictor for the mental rotation score. This was done because a higher probability that the person was male should lead to a stereotype lift for men and a stereotype threat for women. In a second step and in line with Hausmann et al. (2009) and Halpern and Tan (2001), self-ratings on the all items were measured using a 7-point scale from 1 (not at all descriptive of me) to 7 (highly descriptive of me). Cronbach's Alpha for the eight questions was .61. By removing item eight, Cronbach's Alpha was .65. Then, we calculated the mean-score on the seven items related to spatial abilities for each participant.

8.2.2.4 Explicit Evaluative Response

For the explicit rating task three cube figures and three pellet figures were chosen. The explicit evaluative rating task consisted of the following two questions: Familiarity: How familiar are you with the object presented on the screen? (1= very familiar, 7= not at all). Liking: How much do you like the object on the screen? (1= very much, 7= not at all). Gender relatedness: Furthermore, a gender related question was asked: Do you think the object is more female or more male? (1= more female, 7= more male). Each question was asked once with an image of a triangle in a short practice trial and afterwards once for each of the six figures. All questions were asked in a random order on each of the figures and were rated on a 7-point Likert-scale. Participants had five seconds to respond to provoke a spontaneous reaction. For each question and figure type the mean score was calculated for each participant. This differs from the preregistration where we indicated to analyze a composite score. We expected the high experience with mental rotation experiments in our participants to increase only familiarity for the cube figures. Thus, compared to the preregistered analysis, familiarity and liking were not merged into one explicit rating but kept separate.

8.2.2.5 Affective Priming Task

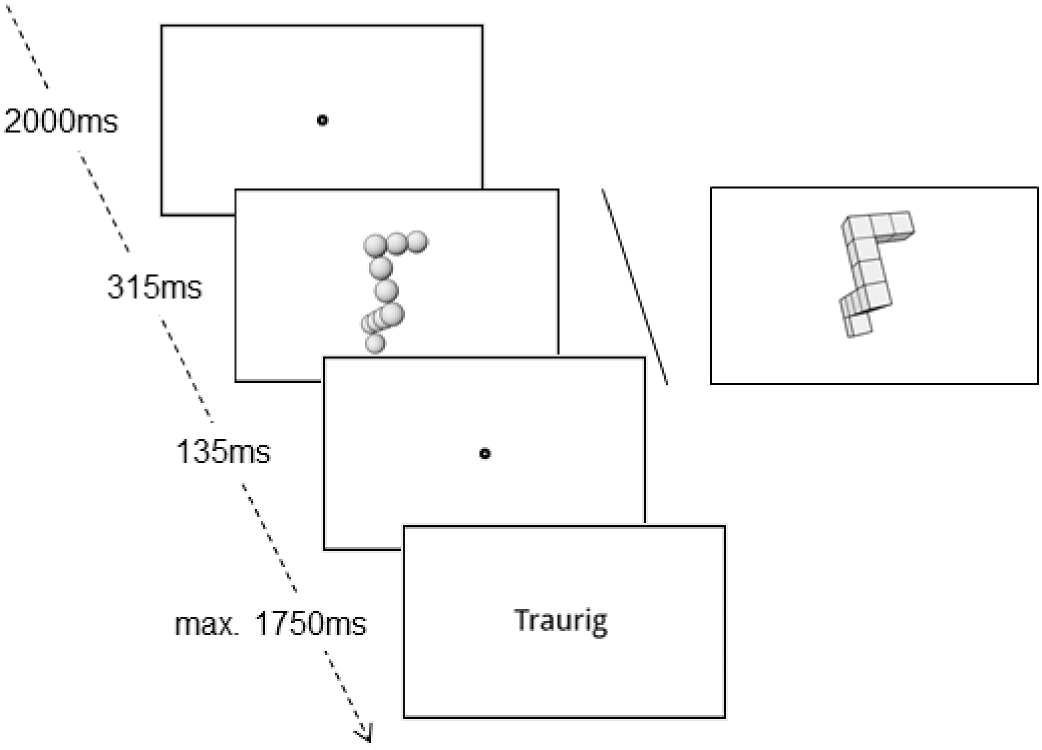
Implicit attitudes were assessed using an affective priming paradigm (Fazio et al., 1995) using ten cube figures and ten pellet figures. We added a short practice trial for the affective priming task with four pictures of a triangle. In the beginning of a trial, a fixation point was shown for 2000ms in the middle of the screen, followed by a picture of either a cube figure or a pellet figure, which was presented for a duration of 315ms. After another 135ms fixation point, a positive or negative word appeared in the middle of the screen, which was chosen randomly from a set of ten negative and ten positive words retrieved from the Berlin Affective Word List (Võ et al., 2009). The participants had to decide as quickly as possible if the word was positive or negative using the arrow keys. The word was presented for a maximum of 1750ms. If the participant failed to respond in this time window, the word disappeared, and the trial was repeated in the end (Fig. 26).

Only those items were analyzed where the correct answer was given. The difference between the mean reaction time of the negative and positive words (negative-positive) was used as indicator for the implicit judgment of pictures of cube and pellet figures (Hutcherson et al., 2008).

Hence, a higher difference score demonstrated a more positive evaluation. Cronbach's alpha for the positive and negative rating of the cube figures was .75 and .80, Cronbach's alpha for the positive and negative rating of the pellet figures was .78 and .74.

Figure 26

Experimental Set-Up for the Implicit Affective Evaluation Showing a Pellet or Cube Figure Before the Word "Traurig" (Sad in German).



8.2.2.6 Mental Rotation Test

Mental rotation performance was assessed using a computerized psychometric mental rotation test (Vandenberg & Kuse, 1978) using two blocks in random order. One block used 12 cube figures and one block used 12 pellet figures (Ruthsatz et al., 2014). Only the items rotated in depth were used. In this test, one target item is presented on the left side and four comparison figures are presented on the right side of the screen. Two of the comparison figures can be rotated into congruence with the target while the other two are mirrored versions. Participants were tasked to identify the congruent items and mark them using the mouse. One point was awarded if and only if both congruent items were marked correctly. Individual trials had no time limit, but each block had a time limit of 6 minutes. In this time and after excluding outliers, participants attempted on average 23.57 (*SD* =1.40) out of 24 trials. Cronbach's alpha for the test with the cube figures was .85 and for the test with the pellet figures .83.

8.2.3 Procedure

The experiment was implemented using the programs OpenSesame (Mathôt et al., 2012) and SurveyJS and it was implemented online on JATOS (Lange et al., 2015). It lasted about 20-30 minutes. In the beginning, demographical data, spatial toy use in childhood, and stereotypes were surveyed. Subsequently, the explicit and the implicit rating tasks and the mental rotation test were conducted, all tasks following the order in which they were mentioned in this section.

8.2.4 Statistical Analysis

Outliers occurred in the affective priming task if a participant gave no answers with the targeted valence for at least one figure type and valence or if a participant gave no answers at all (ten participants). For the mental rotation test, low performance and propensity for guessing was indicated by performance at or below 1/6 correct trials of the attempted trials, that is, those trials in which two items were marked. Two guessers were excluded based on overall performance. Differing from the preregistration, eight further participants scores were excluded from analysis because they did not give two answers on at least one trial for at least one figure type (six participants) or because they selected only one item in more than half of the trials (two participants). Four participants produced outliers for both the affective priming task and the mental rotation test. All outliers were excluded from the respective analyses.

To test if there is a difference between the implicit affective ratings of cube figures and pellet figures depending on the gender of the participant, one 2*2 ANOVA was conducted for the reaction time difference score between negative and positive words with the within factor "type of stimuli" (cube, pellet) and the between-subjects factor "gender" of the participants (men, women). To test if there is a difference between the explicit rating of cube figures and the pellet figures depending on the gender of the participant, a 2*2 ANOVA for each of the three questions (familiarity, likeness, gender relatedness) has been conducted with the within-subject factor "type of stimuli" (cube, pellet) and the between-subject factor "gender" of the participant (men, women). The correlation between explicit and implicit measurements for the two types of stimuli was analyzed separately. This was not explicitly mentioned in the preregistration. After this, a 2*2 ANOVA was calculated for the mental rotation score of the cube and pellet figures the within factor "type of stimuli" (cube, pellet) and the between-subjects factor "gender" of the participants (men, women). In an exploratory analysis a possible gender effect for the psychological variables of spatial toy use, and rating of the own spatial abilities was calculated with a Univariate analysis of Variance and the factor "gender". Equality of variances was tested with the Levene-test and given (all *p*s > .36). For the eight spatial stereotyped probabilities, a MANOVA with the between-subjects factor "gender" was calculated.

For the main research question of relation of implicit and explicit affective response and the mental rotation performance, a Pearson correlation between the psychological measurements (spatial toy use, rating of own spatial abilities, three aspects of explicit evaluations (familiarity, likeness, gender relatedness) and explicit and implicit affective evaluations of the figures) and mental rotation performance was calculated separately for the two figure types. Furthermore, a correlation between the bias towards the own gender identity for the eight spatial items and the performance in the mental rotation test with cubed and pellet figures was calculated. Based on these results, linear regression (method: ENTER) analyses have been conducted.

8.3 Results

8.3.1 Implicit Affective Evaluation

Regarding the reaction time difference in the affective priming task, there was neither a main effect of "type of stimuli", $F(1, 102) = .164$, $p = .686$, $\eta_p^2=.002$, "gender", $F(1, 102) = 1.737$, $p = .190$, $\eta_p^2=.017$ nor an interaction between both factors, $F(1, 102) = .006$, $p = .940$, $\eta_p^2<.001$.

8.3.2 Explicit Affective Evaluation

8.3.2.1 Familiarity

Regarding the score in the explicit familiarity measurement, there was a main effect of "type of stimuli", *F*(1, 112) = 7.946, p = .006, η_p^2=.066, and "gender", *F*(1, 112) = 7.410, p = .008, η_p^2=.062, but no interaction between both factors, *F*(1, 112) = 2.377, p = .126, η_p^2 = .021. Cube figures (*M* = 3.12, *SD* = 1.43) were more familiar than pellet figures (*M* = 3.63, *SD* = 1.37). Also, men (*M* = 3.77, *SD* = 1.15) showed a lower familiarity than women (*M* = 3.17, *SD* = 1.12).

8.3.2.2 Liking

Regarding the score in the explicit likeness measurement, there was no main effect of "type of stimuli", *F*(1, 112) = .071, p = .79, η_p^2=.001, nor "gender", *F*(1, 112) = .007, p = .935, η_p^2 < .001, but an interaction between both factors, *F*(1, 112) = 4.965, p = .028, η_p^2 = .042. Men liked cube figures more than pellets, whereas this is inverted for women, see Figure 27.

8.3.2.3 Gender Relatedness

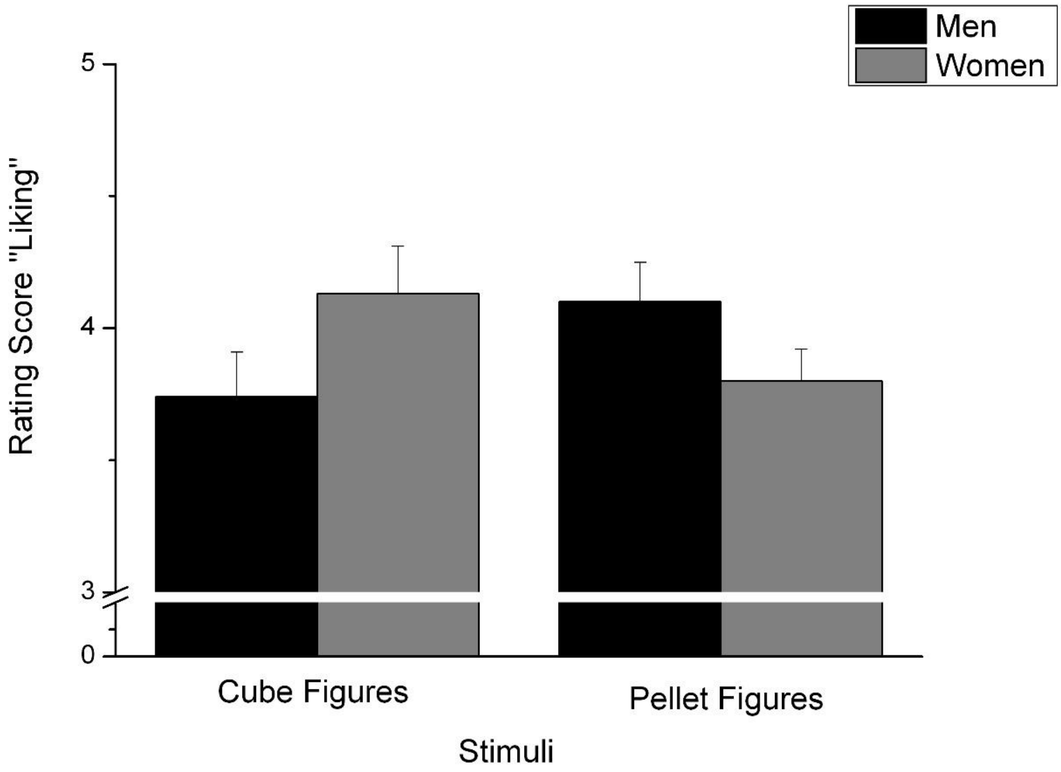
Regarding the score in the explicit gender relatedness measurement, there was only a main effect of "type of stimuli", *F*(1, 112) = 51.49, p < .001, η_p^2=.315, but not for "gender", *F*(1, 112) = .345, p = .553, η_p^2 = .003, and no interaction between both factors, *F*(1, 112) = 2.107, p = .149, η_p^2 = .018. Cube figures were rated more male stereotyped (*M* = 4.73, *SD* = 1.02) than pellet figures (*M* = 3.38, *SD* = 1.18).

8.3.3 Correlation Between Explicit and Implicit Measurements

There was only a small negative correlation between the familiarity of the explicit rating and the implicit measurement of cube figures (r = -.195, p = .048, both other p >. 5) indicating that if cube figures are more known they were rated more positive. There were no significant correlations between the three explicit measurements and the implicit measurement of pellet figures (all ps >. 09).

Figure 27

Explicit Rating Score "Liking" (Mean, Standard Error) Dependent on Type of Stimuli and Separated for Women and Men.

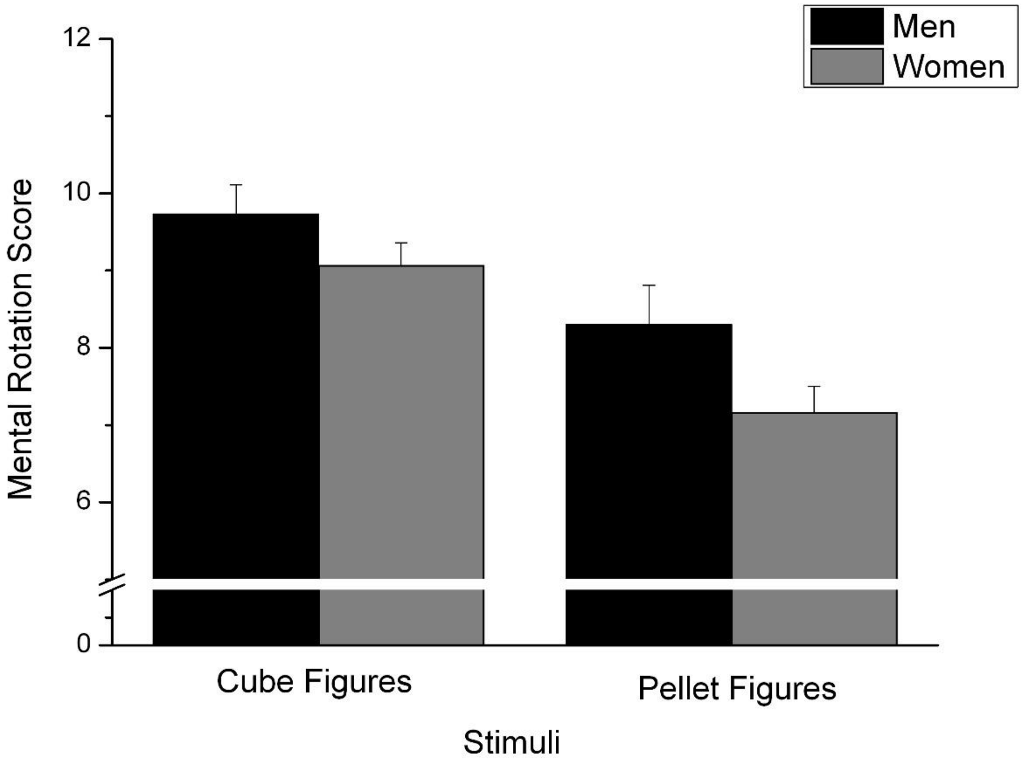


8.3.4 Mental Rotation Performance

Regarding the score in the mental rotation test, there was only a main effect of "type of stimuli", *F*(1, 102) = 58,367, $p < .001$, η_p^2=.364, but not for "gender", *F*(1, 102) = 3.22, $p = .076$, $\eta_p^2 = .031$, and no interaction between both factors, *F*(1, 102) = 1.130, $p = .290$, $\eta_p^2 = .011$. The score for cubes figures (*M* = 9.30, *SD* = 2.42) was higher than the one for the pellet figures (*M* = 7.57, *SD* = 2.94), see Figure 28.

Figure 28

Mental Rotation Score (Mean, Standard Error) Dependent on Type of Stimuli and Separated for Women and Men.



8.3.5 Gender Differences in Spatial Toys Used, Stereotyping, and Spatial Self-Rating

The analyses revealed a significant gender difference in spatial toys use, $F(1, 112) = 43.77$, $p < .001$, η_p^2=.281, but not in spatial self-rating, $F(1, 112) = 2.29$, $p = .133$, η_p^2=.020. Men (M = 0.57, SD = 0.05) used spatial toys more often than women (M = 0.54, SD = 0.06). Furthermore, the multivariate analysis of the probabilities that the person is male for the eight spatial items using Pillai's trace showed no significant effect of gender, $F(8, 97) = 1.025$, $p = .423$, $\eta_p^2 = .078$.

Table 21

Correlation Between the Psychological Variables, the Affective Evaluations (Explicit and Implicit), and the Mental Rotation Scores.

	Spatial toys	Self-rating	Explicit:	Explicit:	Explicit:	Implicit
			Liking	Familiarity	Gender related	evaluation
Cube Figures	r =.093	r =.309	r =-.069	r =.001	r =.071	r =.226
	p =.349	p =.001	p = .485	p =.993	p =.472	p =.025
Pellet Figures	r =.188	r =.298	r =-.045	r =.068	r =--.087	r =.000
	p =.057	p =.002	p =.647	p =.495	p = .382	p =.996

Note. For the explicit ratings, the ratings of the cube figures were correlated with the performance in the test with cube figures, as were the explicit ratings of the pellet figures correlated with the performance in the test with pellet figures.

8.3.6 Correlation and Regression Analysis

All main correlations in relation to the mental rotation scores are presented in table 21. None of the correlations between the bias towards the own gender identity for each spatial item and the mental rotation score with cube figures and pellets were significant (all ps > .075) and not included in the table 1.

The correlation analysis showed that the mental rotation score for the cube figures is correlated with the self-rating (r = .301, p = .001) and the implicit measurement of cube figures (r = .226, p = .025). The mental rotation score for the pellet figures is only correlated with the self-rating (r = .298, p = .002).

The results of the regression with the score for cube figures as criterion and the predictors spatial self-rating and implicit affective evaluation indicated that the two predictors predicted 17.1% of the variance (corrected R^2 = .154, F (2, 95) = 9.814, p < .001). The mean self-rating (*ß* = .371, p < .001) and the implicit affective evaluation (*ß* = .253, p = .013) significantly predicted the mental rotation score for cube figures. The results of the regression with the score for pellet figures as criterion and the predictor self-rating indicated that self-rating predicted 8.9% of the variance (corrected R^2 = .080, *F* (1, 103) = 9.934, *p* = .002). The mean self-rating (*ß* = .298, *p* = .002) significantly predicted the mental rotation score for pellet figures.

8.4 Discussion

In line with our first hypothesis, men liked cube figures more and women liked pellet figures explicitly more, but this did not translate to a significant implicit difference. Cube figures were explicitly rated as more familiar and as more male stereotyped and pellet figures were rated as more female stereotyped by all participants. Contrary to the second hypothesis, the difference in mental rotation performance favoring men and the interaction with the type of figures were not significant, but scores on cube figures were significantly higher than on pellet figures. However and in accordance with our third hypothesis, the implicit affective evaluation positively predicted the mental rotation performance with cube figures. From the investigated possible influencing factors, the self-rating of the own spatial ability predicted the performance in the mental rotation task both with cube and with pellet figures.

8.4.1 Explicit and Implicit Affective Evaluations and Mental Rotation

Only the mental rotation performance of cube figures but not of pellet figures is predicted positively by the implicit affective evaluation. This is in line with the study of Mammarella (2011), who showed that unconscious emotions predict mental rotation performance. In her study, subliminal faces with a neutral, a happy, or a sad expression were presented before solving a mental rotation test with cube figures. Her results demonstrate an influence of the emotional expression in general but do not differ between happy and sad faces. Our study adds that not unconscious emotion in general but unconscious emotional evaluation with respect to the cube figures influence the performance. Participants who have an implicit affective positive attitude towards cube figures show a better mental rotation performance. This result did not appear with pellet figures. Possible explanations might be that cube figures trigger a higher general affective response because they are more familiar than the pellet figures. Another explanation might be related to the study of Palumbo et al. (2015), where the authors not only investigated the positive vs. negative dimension of the perception of curved vs. angular shapes but also their approach and avoidance reaction. Maybe cube figures implicitly trigger an avoidance reaction which hinders the successful solving of a mental rotation task. This idea might be investigated further with the Stimulus Response Compatibility (SRC) task.

For the explicit affective evaluation of the cube figures the results show a gender related answer: Men liked the cube figures more, which were more male stereotyped than pellet figures, and women liked the pellet figures more, which were more female stereotyped than cube figures. This explicit affective evaluation was not related to the performance and did not correlate with the implicit evaluation, which contradicts the explanation of performance differences by gender with the gender stereotype of the stimulus material. Other effects such as the implicit attitude or the difficulty of the stimulus material might be the reason for varying gender differences in the mental rotation performance with other objects.

The study adds to the study of Guizzo et al. (2019) who carved out the relevance of implicit gender spatial stereotyping, at least for men. However, implicit processes are triggered automatically without effort (see Gyurak et al., 2011). In sum, our study shows that implicit affective evaluation, which remain inaccessible explicitly, can affect mental rotation performance. This result is independent of the gender of the participants. There was no gender effect for the implicit evaluation of cube figures and no gender effect in the mental rotation task. The results of this study are in line with the claim of Gyurak et al. (2011) that among others the investigation of implicit emotional processes is a promising research direction in spatial cognition research. An implicitly positive affective attitude towards the objects in a cognitive task might enhance the cognitive performance. This assumption might be worth investigating with other cognitive tasks, such as memory tasks.

Regarding the stereotyping of spatial abilities, our results did not show differences in the stereotypes between genders and no link to mental rotation performance. The non-difference between genders is in line with the results of Hausmann (2014) and Hausmann et al. (2009), whereas only Hausmann et al. (2009) found a connection between the stereotype and mental rotation performance, at least for women. However, both Hausmann (2014) and Hausmann et al. (2009) used the questionnaire to subconsciously activate gender stereotypes, which enhanced performance differences between genders in both studies. Like the evaluation of the stimulus material and emotions, it seems that mostly the subconscious and involuntary stereotyping of spatial abilities is related to mental rotation performance.

It must be noted that the internal consistency measured with Cronbach's Alpha of the stereotyping was rather low. This indicates that there is no consistent stereotyping of spatial abilities in general but rather interindividual differences in the perception of stereotypes for different spatial abilities. While this might also be influenced by the test format, Halpern and Tan (2001) computed at least acceptable values of α = .69 and .76 for their male-typical and female-typical subtests using the same format.

8.4.2 Gender Differences in Mental Rotation

We did not find gender differences in mental rotation performance. At the first glance this contrasts with most psychometric mental rotation tests which find a large gender difference favoring men (Voyer et al., 1995). However, this gender gap is reduced when no time limits are applied (Voyer, 2011). While we did use time limits, these were probably too long as participants attempted almost all trials. Moreover, computerized tests also produce smaller gender differences (Monahan et al., 2008). Another factor reducing gender differences is the realism of figures (Fisher et al., 2018), which might be related to familiarity. Fisher et al. (2018) compared the mental rotation performance in a psychometric mental rotation test with cube figures of four different versions: a) with the standard painted cube figures, b) with photographs of these cube figures, c) with three-dimensional form mounted on boards, and d) when participants were allowed to touch the models either blindfolded or not. In none of the four conditions a gender difference was significant but the decrease in gender difference was also accompanied by an overall increase in performance in three conditions (b, c, and d). Nevertheless, the non-differences between genders are surprising as for the same stimulus material as in this study and in contrast to our nonsignificant performance differences by gender, Rahe et al. (2020) and Rahe and Quaiser-Pohl (2020) found a significant gender difference for both types of stimuli, albeit at much lower average scores. The most important factor leading to the reduced gender differences could thus be the previous experience with mental rotation and ceiling effects.

Our results indicate that cube figures are seen as more male stereotyped than pellet figures, which is in accordance with the study of Palumbo et al. (2015). However, this stereotyping is independent of the gender of the participating student and unrelated to the mental rotation performance, which might be one reason for the missing gender difference as well as the above-mentioned good performance of the participants.

8.4.3 Social Factors and Mental Rotation Performance

Our results demonstrate that the self-evaluating of the spatial ability predict the performance in the mental rotation task. This means that the student participants were able to estimate their own performance. This is an interesting point because self-ratings are important for many things. If one thinks that one is not mathematical talented, one might hesitate to choose a STEM-subject at university. Ackerman and Wolman (2007) demonstrated a moderate to high correlation for the self-rating of spatial abilities and the performance in four spatial tests, paper folding, spatial orientation, spatial analogies, and verbal test of spatial abilities, but no cross-correlation for example to objective verbal abilities. The results are also in line with a study of Quaiser-Pohl and Lehmann (2002) with students of different subjects. Self-ratings of everyday of spatial abilities were the best predictor and accounted for 29.2% of the variance in the performance in a mental rotation test with cube figures. Furthermore, it has also been demonstrated that self-reported navigational ability was significantly associated with most of the spatial memory measurements on a naturalistic Internet-based assessment of spatial memory for environments which had been learned long ago (Selarka et al., 2019).

If there is a relation between self-rating and spatial abilities, it must be investigated if a manipulation of self-rating can enhance spatial abilities. This is an important question because spatial abilities are a crucial component for mathematical abilities (Xie et al., 2020) and are related to STEM education (Moè et al., 2018). In general, self-estimates are strong predictors of professional interests (Neubauer & Hofer, 2020). Thus, the amelioration of the self-rating of spatial abilities could improve the interest in spatial interest areas, such as STEM education or geoinformation.

Our data showed significant gender differences in the use of spatial toys. This is in line with the study of Moè et al. (2018) demonstrating that non-STEM women (as in this study all participants studied Applied Movement Science) preferred spatial toys less than men. Furthermore, the result is in line with the study of Voyer et al. (2000) also demonstrating that women preferred spatial toys less than men. However, in contrast to their study there was no relation between the use of spatial toys and the performance in the mental rotation test neither with cube nor pellet figures. This result must be investigated in more detail with respect to the different groups of participants (students of psychology vs. movement science) and the two different countries, where data were retrieved, but also with respect to the different procedure in the mental rotation tasks. In the study of Voyer et al. (2000) 24 tasks with cube figures had to be solved in ten minutes while in our study 24 tasks with cube and pellet figures had to be solved in six minutes per block (in total 12 minutes). This longer working time in the study presented here might contribute to the overall better performance in comparison to the results of the study of Voyer et al. (2000). This good performance might also be one reason that the gender difference in mental rotation performance was not significant in the study presented here.

8.5 Limitations

One major limitation of the study is the unexpected non-significance of performance differences in mental rotation by gender. The assumptions regarding differences in implicit and explicit rating of the figures were hypothesized to be related to performance differences. The significance and non-significance of implicit and explicit rating differences might not be transferable to participant samples, in which performance differences by gender are larger. This could be related to the high percentage of participants with previous experience with mental rotation tests, which could also be a reason for differences in familiarity and other measures between stimuli. However, the regression is not limited by this as it gathers insight on an individual level.

The fixed choice of the overall order of tests could also influence the results. As evidenced by Hausmann et al. (2009) the stereotype-test can be used as a manipulation to activate gender stereotypes and thus influence mental rotation performance.

The study is also limited by the unequal number of men and women participating. Furthermore, the study has been implemented online. However, the interesting variables of the implicit affective evaluations and the mental rotation performance show a sufficient and good reliability. We chose as an implicit measurement an affective priming paradigm because we wanted to investigate the relation of automatically activated emotional evaluations on the mental rotation performance. The IAT would be the test to choose if the cognitive implicit association is more relevant. Furthermore, the relation of unconscious emotion and the unconscious emotional evaluation should be investigated in more detail. Last, the correlational design of implicit affective attitudes and mental rotation performance does not allow causal conclusions.

8.6 Conclusion

This study shows a positive correlation between the affective implicit evaluation of cube figures and the performance in the mental rotation test with cube figures and a positive prediction of the performance in the mental rotation test with cube figures by this implicit affective evaluation. This relation did not appear with pellet figures and was independent of the gender of the participants. For this, the study contributes to the research on unconscious emotional evaluations and imagery process.

9 Sixth Study: Sex Differences and Trial Design

As mentioned before, sex differences in mental rotation performance are an interesting topic. In line with much research on sex differences, we turned to the psychometric test in the fifth study but neglected the question, why sex differences are much larger in this test compared with the chronometric test. As we also found no sex differences in the fourth study for the design of the second study, we turned to a systematic comparison of trial and test designs to determine whether and where sex differences are affected by them.

9.1 Goal of the Study

The main goal of this study is to investigate the reasons for varying sex differences between different types of mental rotation tests. All mental rotation tests are suggested to measure the same mental rotation ability and the results of different mental rotation tests do correlate with each other (Voyer et al., 2006). However, large performance differences favoring men are only detected in psychometric tests (VK tests) whereas there are only small or no differences in chronometric tests (SM tests) between men and women (Peters & Battista, 2008). For the mental rotation test suggested in the second study (JJ test), the fourth study also indicated no sex differences in performance.

As the reasons for varying sex differences between tests are still inconclusive, two main questions remain:

1) What are the differences between the tests? Can we explain both as specific cases of a general test?
2) Why do sex differences occur mostly in one test (psychometric) but not the other (chronometric)? Which difference(s) is/are responsible?

The following investigation of differences between tests will be organized in two parts. The first part will aim to analyze the properties of mental rotation test designs and provide an overview on the current state of research on their influence on sex differences in performance as well as some possible future directions. The second part will follow up with an experimental manipulation of some of the most promising design properties identified in the first part and their effect on sex differences in performance.

9.2 Part 1: Differences Between Mental Rotation Tests and Their Influence on Sex Differences in Performance

The differences (and similarities) between mental rotation tests can be grouped into two parts. First, there are differences between the individual trials of each test. Second, there are overarching differences of the test design and organization mostly unrelated to individual trials.

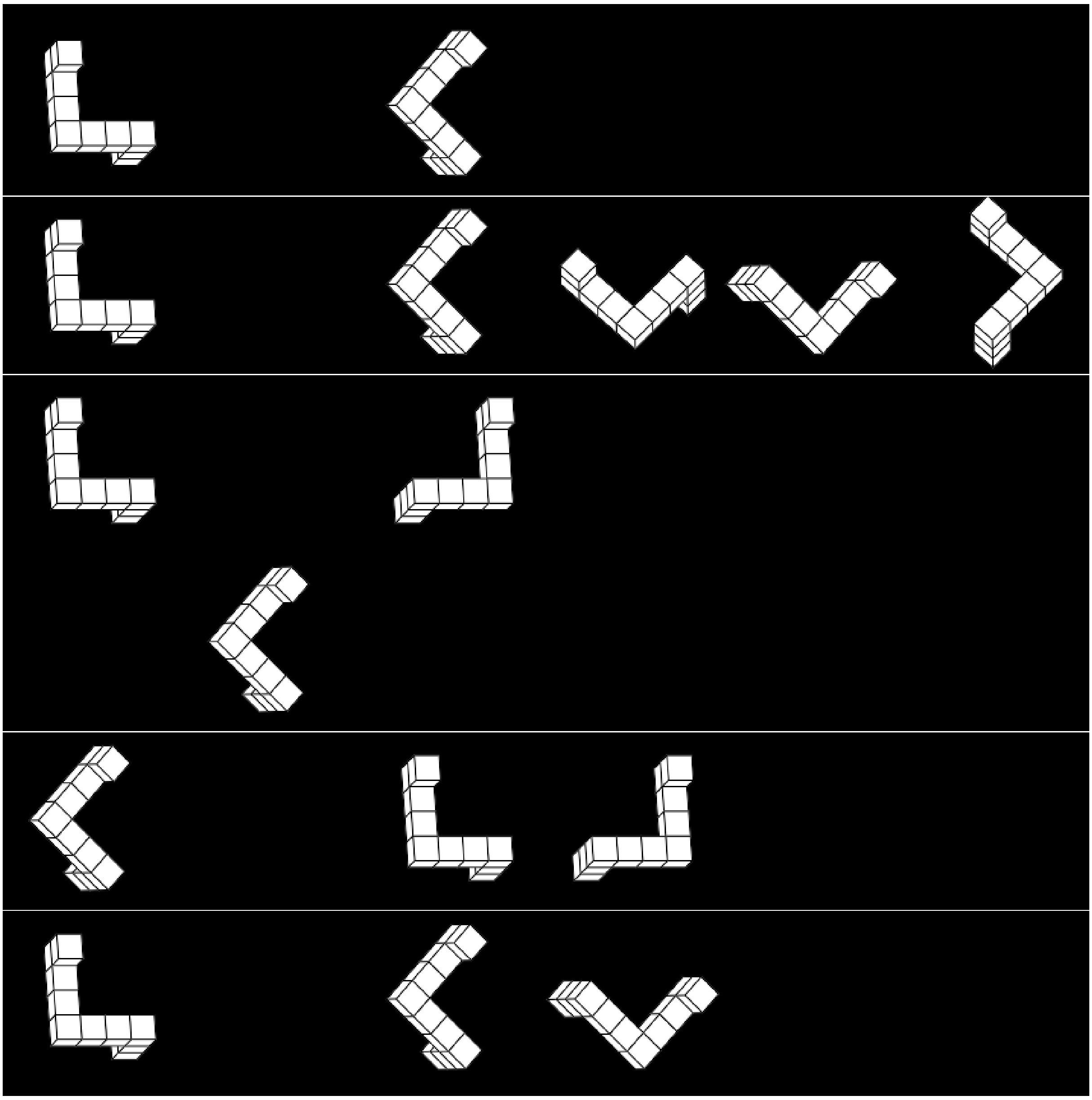
9.2.1 Mental Rotation Trials

9.2.1.1 Differences Between Individual Trials of Different Mental Rotation Tests

While the recent JJ test was inspired by the SM tests, the trials show a similarity to trials of VK tests, as always half of the answers are correct. In that way, it can be seen as a computerized type of VK trials with only two alternatives. Moreover, the two alternatives are aligned across their mirroring plane such that they are easily identifiable as mirrored. In VK trials the four alternatives are not easily pairwise identifiable as mirrored to each other. By modifying the layout of JJ trials to a horizontal positioning of all three figures, the differences and similarities between the tests are shown in figure 29.

Figure 29

Different Mental Rotation Trials.



Note. Exemplary trials from different mental rotation tests. From top to bottom: (1) SM trial. (2) VK trial. (3) JJ trial. (4) JJ trial in horizontal layout. (5) JJ trial without paired choices/VK trial with only two choices. The respective tasks are to decide whether the two figures are the same or different (1) or to select exactly half of the alternatives, which are rotated versions of the separated target item to the left (2, 4, and 5) or bottom (3).

Thus, we propose three major differences between trials of the three mentioned mental rotation tests:

1) Compared with SM trials, participants are informed that always exactly half of the possible answers are correct in both JJ trials and VK trials. In most versions of the SM test, overall, about half of the trials are "same" trials and the other half "different" trials but this is typically neither exact nor known to participants.
2) The trials use different number of alternatives (1, 2, or 4) and total number of items per trial (2, 3, or 5).
3) In JJ trials, the alternatives are pairwise identifiable as mirrored and aligned across the mirroring plane whereas in VK trials, they are ordered randomly and not pairwise mirrored. For the SM trials, this distinction is not applicable.

Additionally, there is a possible difference in the upright orientation of the figures. For abstract figures such as the widely used cube figures and polygons, this is somewhat arbitrary as there is no clear upright orientation. While we can for the cube figures describe the upright orientation as an alignment with the natural axes (which are most of the time also the rotational axes), this is often not reported and might need further investigation.

While the differences in the layout were necessary to highlight the differences between tests, they do not seem to play a major role in test performance. For SM trials, Xue et al. (2017) found no difference whether the left or the right item was rotated. For a version of VK trials with a vertically separated target, J. K. Krüger and Suchan (2016) found large sex differences in their control group, which was a sample of the general population. A comparison of horizontal distances between stimuli mentioned by Battista and Peters (2010) also revealed no differences. Layout variations have been used without further inspection of the layout itself by Wohlschläger and Wohlschläger (1998), Wohlschläger (2001), and M. Krüger and Krist (2009) and different experiments regularly use varying computer screens and image sizes as well as non-standardized distances to the screen or test material and resulting optical distances between stimuli. Sex differences in VK tests and non-differences in SM tests have emerged under various of these non-systematically varied conditions.

Out of the suspected major differences, only the number of items per trial has yet been manipulated. Both Titze et al. (2008) and Voyer et al. (2020) restricted the visibility of the answer alternatives in VK tests. In the study of Titze et al., three out of four alternatives were hidden by a template and participants only saw the target and one alternative at a time. In two experiments, Voyer et al. employed a gaze-contingent display and participants could only see one item (either the target or one of the alternatives) at one time. However, these manipulations did not aim to change the number of alternatives and items within each trial, but only the visibility. To actually manipulate the number of items per trial, Titze et al. (2010) conducted a test that resembles a paper and pencil version of a SM test. Each trial of a VK test was split into four pairwise same/different judgements, which were then presented in random order. All three studies reported significantly better performance for men compared with women in all experiments.

9.2.1.2 A General Mental Rotation Trial

Whereas the previous chapter dealt with a comparison of existing mental rotation trial, there is the possibility to further change parameters of the trial design and get new mental rotation trials. Despite the exploratory nature of such an approach, it could lead to better measurements and understanding of mental rotation ability. While both VK and SM tests are widely used, the choice is somewhat arbitrary, as in someone used these designs and as the results proved interesting everyone else used the same designs. There seems to be no clear rationale why mental rotation tests should be presented in this way and no other except for comparability. In theory, all tests should measure mental rotation ability and thus be comparable. Further exploration could lead to better understanding of design parameters and what mental rotation ability is, where it might end, and where other sex sensitive abilities come into play (and by which parameters they are enhanced).

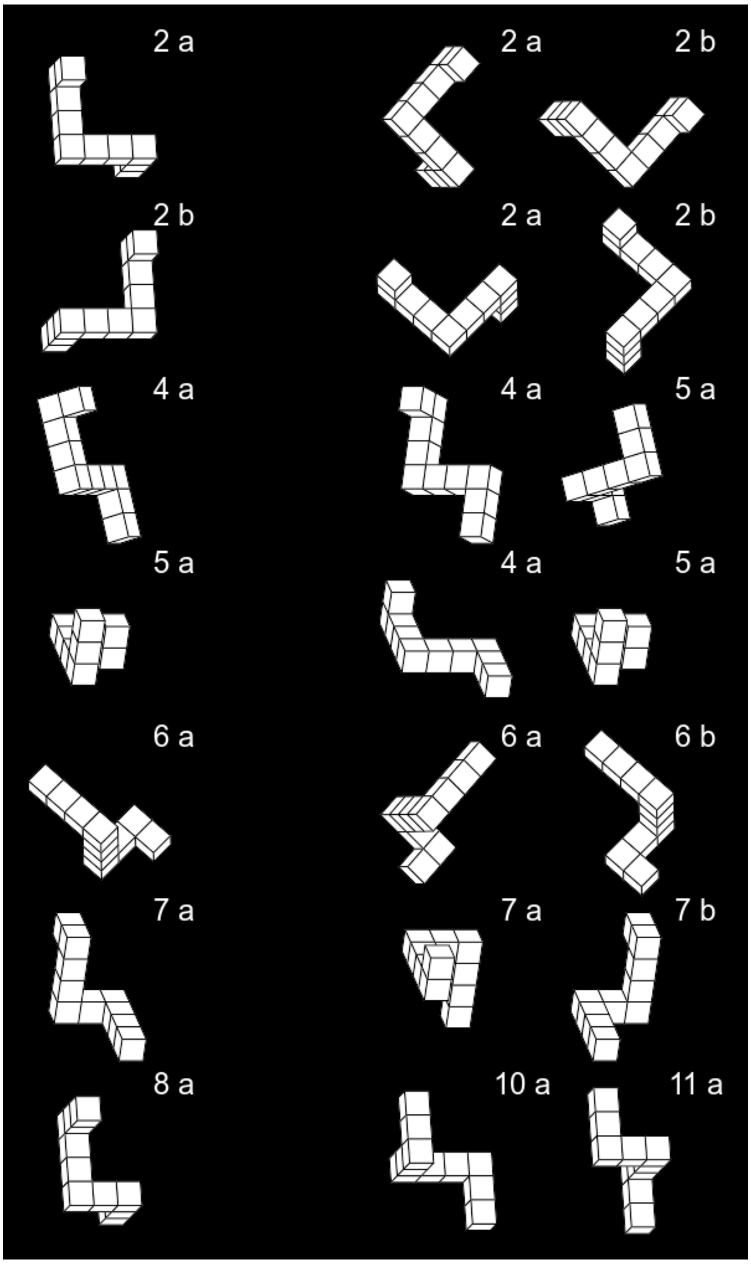
While using the same test is desirable for comparability between studies[10], both the development of the JJ design and the increased possibility of single item analysis by deviating from the 50% correct requirement discussed as part of the following power analysis highlight possible improvements in test designs, which in turn could lead to better detection of effects in question. Boone and Hegarty (2017) also already suggested to modify VK tests by removing trials that allow for nonrotational strategies to get a more pure measurement of mental rotation ability. Thus, further exploring different trial designs could prove fruitful for testing mental rotation ability and separating it from interfering abilities incorporated in the solution of existing tests.

The manipulation of mental rotation trials could explore extension and changes in further parameters that were not yet considered. This could include changing the number of correct answers and/or providing other information about the number of correct answers (e.g., at least one and at most three out of four are correct or no information about the number of correct answers). With a deviation from the requirement that always 50% of the alternatives are correct, odd numbers of alternatives are also possible. For now, all tests also show only one target. In the second study, we proposed that VK tests could also be modified by a second mirrored target to the right of the alternatives. This can also be extended to more targets if structurally different figures are used.

[10] While many studies use the SM or VK tests, there is great heterogeneity in the details as reviewed in this and the following chapter.

Figure 30

Example of a General Mental Rotation Trial.



Note. Each model is labeled by its number and orientation (a or b) for easier identification. The task would be to identify all matching alternatives (right columns) for each target (left column). This example includes varying numbers of correct alternatives (zero, one, or two) for different targets, both mirrored (6b, 7b) and structural (10a, 11a) distractors without matching targets, and also mirrored targets (2a, 2b). Information about the number of matching alternatives per target could optionally be provided.

For that, it is possible to propose a general mental rotation trial. Such a design could employ any number of targets and alternatives with the task to find all rotated alternatives for each target. There can be varying number of alternatives to each target (including 0) and alternatives with no congruent target (as is the case with distractors in existing VK tests and mirrored trials in SM tests). Information about the numbers of correct alternatives can be provided to participants with varying precision. One such possible mental rotation trial is shown in figure 30. While it does incorporate all existing mental rotation tests as special cases, there is no claim for completeness.

9.2.2 Differences Between Mental Rotation Tests

Next to the differences between individual trials of the tests, there are some overall differences between the tests. These were already partially manipulated by researchers, although not always as part of an analysis of sex differences. In the following, we describe some of these manipulations and their influence on sex differences. These studies do however often vary in more than the parameter in question (e.g., different time limits for digital tests compared to paper and pencil tests) and use different numbers of trials (we expect smaller standardized effect sizes of sex differences with less trials due to the increased relative variance of guessing).

9.2.2.1 Ending Condition/Time Limits

In their most often used versions, the VK test is limited by both time and the number of trials (two times 3 minutes for 12 trials, Peters et al., 1995), the SM test is limited by the number of trials, and the JJ test is limited only by time. By relaxing or removing time limits in VK tests, sex differences in performance were reduced, but still easily detectable at Cohen's $d \approx 0.5$ and much larger than in SM tests (Peters, 2005; Voyer, 2011). As sex differences are neither commonly observed for SM tests limited by the number of stimuli and JJ tests limited by time, the ending conditions of tests cannot be the sole reason for sex differences.

9.2.2.2 Stimulus Material

Whereas all tests most commonly use abstract cube figures as stimuli, there are small differences between the used figures. However, the original stimuli of Vandenberg and Kuse (1978) were based on the stimuli developed by R.N. Shepard and Metzler (1971) and redrawn versions of the stimuli have been used interchangeably between tests. For example J.K. Krüger and Suchan (2016) used the figures of Peters and Battista (2008), which are widely used in SM tests. Thus, other aspects of the stimuli and different stimuli have been investigated as possible reasons for sex differences in VK tests. The possible reasons include the embodiment of figures, gender stereotypes about the items, and the realism of figures. Moreover, the occlusions in the two-dimensional representations of three-dimensional objects have been analyzed.

For human figures and a separation of partially occluded and nonoccluded figures, sex differences are still reported in VK tests, although often reduced for tests with overall higher average scores (e.g. Alexander & Evardone, 2008; Doyle & Voyer, 2013; Jansen & Lehmann, 2013; Voyer et al., 2020). As the widely used cube figures are seen as more male stereotyped, Ruthsatz et al. (2014, 2015, 2017) investigated female stereotyped stimuli and partially found reduced and even reversed sex differences in the VK test performance of children. These results have however only been partially transferred to adults with reduced or negated sex differences but no female advantage (Rahe et al., 2020; Rahe & Quaiser-Pohl, 2020; see also the fifth study). Regarding the realism of stimuli, both Fisher et al. (2018) and Robert and Chevrier (2003) found no significant sex differences for real three-dimensional objects resembling cube figures. However, men were much faster than women in the study of Robert and Chevrier and scores approached the maximally attainable scores in the study of Fisher et al., indicating possible ceiling effects. While it is possible that more realistic figures reduce sex differences, no clear conclusions can be drawn as neither study was sufficiently powered to reliably detect smaller than large effects with only about 20 men and women per condition.

For SM tests, Voyer and Jansen (2016) did find better performance for men with human figures as stimuli[11]. However, this is one of few studies also reporting better performance for men for cube figures and the performance differences between stimuli were comparable for all stimulus types or even larger for the human figures (although quantified by differences in reaction time and accuracy on both same and different trials). Using hands as stimuli, Voyer et al. (2017) found better performance for men, whereas Campbell et al. (2018) found possibly better performance for women. By comparing multiple stimulus types in SM tests, Jansen-Osmann and Heil (2007b) found sex differences only for two dimensional polygons. Heil and Jansen-Osmann (2008) replicated these sex differences for polygons and found larger differences for more complex polygons. Except for the polygons, all reported standardized effect sizes for sex differences in SM tests are small. As the cube figures show the largest reaction times and lowest accuracies in comparisons of different stimuli and thus are the most complex in the study of Jansen-Osmann and Heil (2007) there does not seem to be a monotonous relationship between complexity and sex differences in SM tests. However, it is not clear if the same effect occurs for VK tests as polygons have not yet been used as stimuli. Overall, it is likely that the used stimulus material and its complexity interacts with sex differences in mental rotation tests but that they are not a major reason for differences between different tests using cube figures.

9.2.2.3 The Rotational Axis

For many versions of VK tests only rotations in depth are used, whereas SM tests often use rotations in the picture plane or combine items rotated both in depth and in the picture plane in one test. In a study of a VK test with children, Ruthsatz et al. (2014) found an interaction of gender and axis: Larger sex differences occurred for cube figures rotated in depth compared with rotations in the picture plane. This effect was however not found for the second stimulus type, pellet figures. For adults, Battista and Peters (2010) did find better performance for rotations

[11] Similar to different versions of cube figures, different versions of human figures have been used between different tests.

around a vertical axis in depth compared with a horizontal axis but no significant interaction with sex differences.

For SM tests using only rotations in depth, no or only small sex differences are found (Jordan et al., 2002; Peters, 2005; Rahe, Ruthsatz, Schürmann, et al., 2019). The aforementioned sex differences for polygons are also only found for rotations in the picture plane. Many tests use multiple rotational axes as the widely used stimulus library of Peters and Battista (2008) provides stimuli rotated around all three canonical axes. The original study of R. N. Shepard and Metzler (1971) already suggested a strong similarity between rotational axes and these are often not further distinguished. For realistic stimuli, there is some evidence for better performance for rotations around a vertical axis compared with a horizontal rotation in picture plane (Foulkes & Hollifield, 1989). A possible reason could be the relevance of vertical rotations and gravity in real-life but as Shiffrar and Shepard (1991) point out, these differences could be related to how the rotational axes are aligned with the features of the stimuli even for abstract figures. Despite the possibility for much further research into the effect of rotational axes, it seems unlikely that rotations in depth are a major reason for the discrepancy between sex differences found in different tests.

9.2.2.4 Type of Distractors

Both SM tests and JJ tests use mirrored items as distractors whereas VK tests also use structurally different items. By separating the answers for trials with mirrored and structural distractors in VK tests, mostly either non-significant effects of the distractor type or reduced sex differences and overall better performance for structural distractors are found (Boone & Hegarty, 2017; Bors & Vigneau, 2011; Doyle & Voyer, 2013; Monahan et al., 2008; Voyer & Hou, 2006). Thus, it seems that structural distractors in VK tests are likely not the reason for sex differences. However, Boone and Hegarty (2017) also analyzed the incorporation of structural distractors with or without mirrored distractors in SM tests and found large sex differences, indicating that the combination with structural distractors is a possible reason for sex differences. But as they also found large sex differences on a VK test using only mirrored distractors in line with other uses of only mirrored distractors in VK tests (Rahe et al., 2020; Rahe & Quaiser-Pohl, 2020), their results only open this possibility to increase sex differences in SM tests but not how to reduce sex differences in VK tests.

9.2.2.5 Test Administration

The VK tests were originally developed and are still often administered as a paper and pencil test, whereas SM and JJ tests are computerized to measure reaction times of individual trials. With the rising prevalence of computers since the original conception, many computerized and online versions of VK tests have been administered in more recent studies. These tests regularly report better performance for men, although sometimes with reduced effect sizes compared with paper and pencil tests (e.g. Debarnot et al., 2013; Krüger & Suchan, 2016; Monahan et al., 2008; Peters et al., 2007; Voyer et al., 2020). As mentioned before, in one study of a paper and pencil SM test, Titze et al. (2010) did find large sex differences in performance. While the test administration cannot be the sole reason for varying sex differences between mental rotation tests, the results of Titze et al. need to be investigated further.

9.2.2.6 Participant Organization

As with the test administration, the VK test was developed to quickly test large groups and multiple participants often perform the test in the same room at the same time. The group composition and stereotypes associated with genders have been hypothesized to influence performance differences (Moè, 2018). However, associated with the form of the test administration, computerized and online versions of VK tests are often conducted individually or in the same manner as SM tests and still produce sex differences. Moreover, also individual testing using paper and pencil VK tests still produces sex differences (e.g. Titze et al., 2008).

9.2.2.7 Scoring System

To reduce the impact of guessing and based on the original work of Vandenberg and Kuse (1978), the most widely used scoring system for VK tests awards one point to a trial if and only if both correct alternatives are selected. These scores are not directly comparable to SM tests, for which the reaction time and accuracy is analyzed for each "same" item and not analyzed for "different" items. For JJ tests, there is no need to remove items, but reaction time and accuracy are still evaluated on a by-item basis. While participants are mostly instructed about the scoring system in VK tests, the task in SM tests is typically to solve the items as quickly and accurately as possible.

Due to the combination of multiple items into one trial, the reaction time in VK tests cannot be attributed to single items. For a comparison of scoring systems on VK tests, Voyer et al. (1995) found smaller, but still large sex differences for test scores on single items. For a comparison of both scoring systems within one test, Titze et al. (2010) found a high correlation and large sex differences for both. In the few studies analyzing both performance for mirrored stimuli and sex differences in SM tests, Peters (2005) observed no sex differences and Voyer and Jansen (2016) found small and comparable sex differences on both mirrored and non-mirrored trials. Kerkman et al. (2000) did find sex differences only for the mirrored trials, indicating a possible reason for the worse performance of women in VK tests. Their design however was different as very different stimuli were used and participants were also required to identify the direction of rotation. Moreover, their results indicate the tendency of women to guess that trials are non-mirrored (leading to higher accuracy for non-mirrored trials and lower accuracy on mirrored trials). While this led to sex differences only for the mirrored trials, this could also be interpreted as an overall worse performance. While the differences in the scoring systems prevent most results from being directly compared between tests, they may be a possible reason for reduced performance differences between sexes and could use some further investigation.

9.2.2.8 Practice Trials and Feedback

The mental rotation tests also differ in the number of practice trials before results are recorded. While VK tests most commonly use three or four practice trials (composed of 12 to 16 alternatives), there is no standard for SM or JJ tests. The number of practice trials ranges from zero to over 100 with many studies using ten to 20 practice trials comparable with VK tests, but there does not seem to be a systematic investigation of sex differences depending on the amount of practice. Moreover, in repeated VK tests, sex differences are still found in later tests (Peters, 2005; Peters, Laeng, et al., 1995) despite the increased practice, although sometimes reduced (Peters, Chisholm, et al., 1995). Another aspect is the inclusion of feedback for every trial during SM tests in some procedures. Rahe et al. (2019) identified the missing feedback as a possible reason for sex differences. However, for example both Jansen-Osmann and Heil (2007b) and Voyer and Jansen (2016) used feedback for all trials and partially found sex differences.

9.2.3 Possible Influences on Sex Differences

Next to the difference between trials whether the alternatives are pairwise easily identifiable as mirrored to one another, which has not yet been investigated, the further comparison between tests has revealed some possible influences on sex differences. Overall, reducing task difficulty (by using simpler stimuli, easier identifiable distractors, or increasing available time) or digitalizing VK tests reduces sex differences, but sex differences are still observed in VK tests but not or with smaller effect sizes in SM tests independent of these parameters. Moreover, even a combination of multiple factors reducing sex differences does not eliminate them. For example, Doyle and Voyer (2013, 2018) and the free-viewing experiment of Voyer et al. (2020) used computerized VK tests without time limits and human figures as stimuli and found better performance in men. As the effects of task difficulty have not yet been isolated and compared between tests and digitalizing tests typically influences other factors of the test design, there are two issues for which some more discussion seems appropriate. The first, related to the sex differences found by Titze et al. (2010) on a paper and pencil SM test and the possibly reduced sex differences on computerized VK tests, is the number of alternatives and trials presented at one time. The second, related to reduced sex differences on VK tests with overall higher scores and the difficulty of comparing results between tests, is the complexity of single item comparisons.

9.2.3.1 Number of Alternatives per Trial and Simultaneously Presented Trials

In general, all mental rotation tests can be solved by repeated pairwise comparisons but differ in the number of necessary and possible pairwise comparisons to solve each trial. The overall number of possible pairwise comparisons between two figures are: one for SM trials, three for JJ trials, and 10 for VK trials. Out of these, one, two, and four involve the target. Due to the information that exactly half of the alternatives are correct only one comparison for JJ trials and two or three comparisons for VK trials are necessary to complete the trial as the other items can then be solved by exclusion. Between computerized and paper and pencil tests, the tests also differ in the number of visible stimuli unrelated to the trial at hand as multiple trials are presented on one page. These offer further possible comparisons, which are not necessary and not helpful to find the solutions of individual trials. Either the number of necessary comparisons per trial, the number of possible comparisons per trial, the overall number of possible comparisons within a test, or a combination of all of them could be related to sex differences as all offer additional comparisons, which are not related to test performance. However, due to a lack of studies, neither effect nor their interaction has been conclusively investigated.

Related to the comparisons within one trial, both the template task of Titze et al. (2008) and the restricted viewing experiments of Voyer et al. (2020) still required the same number of same/different judgements for one trial as the classical VK trials. However, in the trials of Titze et al. all pairwise comparisons involved the target. Regarding possible comparisons between trials, the paper and pencil SM test of Titze et al. (2010) had multiple and possibly related trials (same target) being visible at once. Next to strategies of optimally disentangling the linked same/different judgements within one trial and the unlinked comparisons between different trials, both the number of comparisons necessary and the number of visible stimuli could be a measure of "perceived complexity". These could possibly be influential for sex differences and thus also explain smaller sex differences in computerized tests.

9.2.3.2 The Difficulty of Individual Pairwise Comparisons

As mentioned, the tests use different scoring systems and individual items cannot be compared. As there seem to be smaller sex differences for less difficult trials in VK tests and less difficult test versions, a higher difficulty of VK tests could be a possible reason and needs further investigation. However, the lower score on VK tests compared with the overall error rates in SM tests must at least be in part due to the scoring system, which makes it more difficult to achieve points. Thus, some approximation of error rates for individual items in VK tests seems necessary.

Due to the information that always half of the alternatives are correct, the individual items are not independent. However, the exact dependency depends on the employed strategy and is impossible to estimate in general. In the following, we present two simple approximations of overall accuracy of single item comparisons in VK tests. These simple methods only estimate overall accuracies and do not account for differences between trials (e.g., lower accuracy with larger angular disparity). While a typical correction for multiple-choice tests concerns the number of answer alternatives (i.e., six possibilities to choose two out of four alternatives in VK tests), this is not necessary for a comparison between SM and VK tests. As we try to compute the probability that one single item in VK tests was solved correctly from the overall score, this includes the probability of guessing correctly in half of the cases. As a result, the corrected scores are directly comparable between tests as the probability to correctly identify rotated pairs, although they do not represent the true accuracy.

One way to estimate the probability p to solve a single item of a VK test correctly is to assume that trials are either solved by knowledge or guessed, that means the probability to solve a trial is either 1 or $1/6$. Thus, $5/6$ of the guessed trials are solved incorrectly[12]. This means, that the number of trials that were guessed correctly are $1/5$ times the trials that were solved incorrectly. By transforming the guessing probability to chance level of ½ for single items, the resulting average probability is thus

$$p_1 = \left(\left(s - \tfrac{1}{5}(n-s)\right) + \tfrac{1}{2}\left(\tfrac{6}{5}(n-s)\right)\right)/n = \left(s + \tfrac{2}{5}(n-s)\right)/n = \tfrac{3s}{5n} + \tfrac{2}{5},$$

[12] This does not account for trials in which one, two, or three items can be solved by knowledge because the resulting probability depends heavily on the strategy.

where s is the achieved score, n is the number of trials, and $n - s$ is the number of trials solved incorrectly.

Another way to estimate single item accuracies is assuming the same probability for each item. This however is heavily dependent on the employed strategy. Assuming that a trial is solved correctly if three single item comparisons are solved correctly[13], the probability to solve a trial correctly is $s/n = p_2^3$, that is $p_2 = \sqrt[3]{s/n}$. For this simple measure, variance in p_2 is neglected. For a given p_2, a variance between items within one trial would lead to lower scores than p_2^3, whereas a variance between trials but not between items would lead to higher scores than p_2^3. Note that $p_2 > p_1$ for the range of interest ($\frac{1}{6} < \frac{s}{n} < 1$) as $p_2^3 - p_1^3 = 0$ for $\frac{s}{n} = \sim - 3.1, \sim 0.1, 1$.

Using either method, the overall accuracy for single task comparisons is estimated as .59 to .95 for various studies with the lower values attained by women in studies employing time limits (see table 22). Assuming three to five single item comparisons are performed per trial (independent of the assumption for p_2 that of those only three are used to find the solution) and time limits of 3 to 6 minutes, these would be achieved at average reaction times of 1.5-5 seconds[14]. Especially the male scores are thus comparable to accuracies in SM tests ranging from .6 at an average reaction time of ~5 seconds (pretest data for the rotations in depth of Adams et al., 2014) to over .96 at average reaction times of ~3 seconds (R. N. Shepard & Metzler, 1971) or "close to 100%" (Peters, 2005) with most studies reporting accuracies of .85 or higher at average reaction times of 2-3 seconds. These typically only include rotated items, whereas reaction times are typically larger on mirrored trials. On the other hand, if it is reported the accuracy on mirrored trials is mostly higher

[13] One possible strategy is to solve items until two same or two different items are found and solving the remaining items by exclusion. If items were chosen randomly, there is a probability of 1/3 that two items have to be solved and a probability of 2/3 that three items have to be solved. However, if items are solved from left to right (or right to left), in the most used version of the test only two out of 24 trials are solved correctly with only two items.

[14] This comparison is independent of the number of finished tasks if one assumes that participants solve the same number of trials over time when trading speed for accuracy, but this does not hold for other possible speed-accuracy trade-offs (Liesefeld & Janczyk, 2019).

and also falls in the range of the estimated male VK performance. Table 23 provides an overview of average reaction times and accuracies for some studies using SM tests.

Table 22

Estimated Single Item Accuracy From VK Test Scores Using the Proposed Calculation Methods.

Study	Score(n)	p_1	p_2
Alexander and Evardone (2008) Block figures	Men 6.74 (12)	.74	.83
	Women 3.85 (12)	.59	.68
	Men ratio 0.71	.83	.89
	Women ratio 0.48	.69	.78
Titze et al. (2008) without pattern	Men 20.76 (24)	.92	.95
	Women 18.91 (24)	.87	.92
Doyle and Voyer (2013)	Men ratio .65-.81	.79-.89	.87-.93
	Women ratio .47-66	.68-.80	.78-87
Monahan et al. (2008) computer test	Men 12.34 (24)	.71	.80
	Women 9.57 (24)	.64	.74
Monahan et al. (2008) paper and pencil test	Men 12.83 (24)	.72	.81
	Women 8.15 (24)	.60	.70
Peters (2005)Study 1, 3 minutes	Men 12.6 (24)	.72	.81
	Women 8.8 (24)	.62	.72
	Men ratio 0.739	.84	.90
	Women ratio 0.602	.76	.84
Peters (2005)Study 2, 6 minutes	Men 18.0 (24)	.85	.91
	Women 14.4 (24)	.76	.84
	Men ratio 0.776	.87	.92
	Women ratio 0.642	.79	.86

Note. The ratios for the study of Peters (2005) were calculated from the reported percentage of problems attempted.

Table 23

Average Reaction Times and Accuracies in SM Tests Using Cube Figures.

Study	Reaction time (s)	Accuracy
Adams et al. (2014) picture plane	4.7	.80
Adams et al. (2014) depth	5.2	.61
Jansen-Osmann and Heil (2007b)	1.84	.954
Jolicœur et al. (1985) experiment 1	4.5	.87
Paschke et al. (2012)	~2.5	~.85
Paschke et al. (2012) Mirrored trials	~2.7	~.95
R.N. Shepard and Metzler (1971)	~3	.968
R.N. Shepard and Metzler (1971) Mirrored trials	3.8	
S. Shepard and Metzler (1988)	2.06	.906
S. Shepard and Metzler (1988) Mirrored trials		.942
Voyer and Jansen (2016)	2.10	.874
Voyer and Jansen (2016) Mirrored trials	2.52	.890
Wiedenbauer et al. (2007) experiment 1	~2.6	.882
Wiedenbauer et al. (2007) experiment 2		.876

Note. For intervention studies, pretest data is shown. For studies that estimate rotation speed, the reaction time is calculated for an angle of 90°. Values marked with ~ are estimated from figures.

Based on these estimations, there do not seem to be major differences in single item difficulty between tests, at least for men. While reducing the overall difficulty of VK tests often reduces sex differences in performance, the non-differences in SM tests can likely not be explained by the difficulty of single items. Moreover, varying stimulus material in SM tests has rarely produced sex differences and these are not systematically related to task difficulty. Accuracy also drops and reaction time increases with larger angular disparity and sex differences have not consistently been observed with this increasing difficulty. This further supports that differences in single item comparisons are not likely to be the main reason for sex differences between VK and SM tests.

9.2.4 Discussion

We have reviewed differences and similarities both overall and between individual trials of different mental rotation tests, which offers possible directions for future research. Because sex differences are often assumed to be in mental rotation ability, the manipulation of VK tests, which produce the largest sex differences, is suspected as a means to uncover the reasons, but the fact that sex differences do not occur or are much smaller on other tests of the same ability is often neglected. However, by comparing the trials of different tests and past research on sex differences, it does seem likely that the trial design is an important factor in the search for sex differences. Whereas much research has focused on reducing sex differences in VK tests, another interesting question is how to increase sex differences in SM tests. Boone and Hegarty (2017) identified the incorporation of structural distractors to achieve this but the removal of structural distractors in VK tests still produced large sex differences. Using polygons as stimuli has produced large sex differences even in SM test performance but they have not yet been investigated for VK tests. It seems possible that polygons as stimuli would produce even larger sex differences in VK tests. Whereas polygons have been varied in complexity, similar approaches have not been employed for cube figures. They have only been compared to other item types despite the multiple possible modifications due to the abstract nature. Next to overall effects of complexity similarly to the polygons, the third dimension also allows the investigation of the alignment with the canonical axes and with the rotational axes. While not directly contributing to reducing sex differences, exploring the effects of the trial design and the possibility to also increase sex differences could help us pinpoint the exact reasons und thus offer further understanding of the occurrence of sex differences in performance.

9.3 Part 2: Experimental Investigation of Within-Trial Factors and Sex Differences in Mental Rotation Performance

In this part, we aim to investigate the influence of the previously identified differences in the trial design on sex differences in test performance. The two most promising parameters seem to be the number of alternatives to each target and whether the alternatives are pairwise easily identifiable as mirrored to one another.

We will use a combination of paired or mixed alternatives with varying numbers of alternatives (two, four, or eight) and the following hypotheses are investigated:

a) We expect larger performance differences favoring men for the test using mixed alternatives.

b) We expect larger performance differences favoring men for tests using more alternatives.

c) Moreover, we expect the effect of the number of alternatives to be larger for the mixed alternatives.

As secondary hypotheses, we will further investigate the following covariates to control variance both within and between subjects. Within subjects the angular disparity of items, the test blocks, and the position of items within blocks will be analyzed. The angular disparity of items is included because we will analyze single items and it is well known from chronometric tests that accuracy decreases with increasing angles (e.g., R. N. Shepard & Metzler, 1971). Blocks will be analyzed due to known learning effects between blocks within tests (e.g., Peters, 2005). The position within blocks is more exploratory as it is affected by less attempted problems for later positions (e.g., Peters, 2005), which decreases accuracy if no distinction is made between attempted and unattempted problems. On the other hand, similar to the learning between blocks and within chronometric tests (see the second, third, and fourth study) we expect better performance for later problems. We will analyze the position within blocks only for the attempted problems and expect an improved performance for later problems.

The experiment was originally planned as a lab experiment but moved online due to the third wave of the COVID-19 pandemic. To account for the larger between-subjects variance in performance due to the wider target population, we will include the covariates education (STEM student, non-STEM student, no student) and previous experience with mental rotation tasks (yes or no). Second to sex differences, education has shown to account for the most variance in mental rotation performance without specific practice (Peters, Laeng, et al., 1995). An interaction of education with sex differences is possible (Moè et al., 2018; Peters, Laeng, et al., 1995). Practice of mental rotation has also shown large increases of performance but the interaction with sex differences is not clear (Peters, Chisholm, et al., 1995; Peters, Laeng, et al., 1995). As practice has shown almost linear improvements, which would necessarily reduce both sex and education differences after extensive practice due to ceiling effects, we will explore the interaction of practice with sex differences and education.

9.3.1 Method

9.3.1.1 Power Analysis

The power analysis was performed using simulated data for estimated scores and effect sizes based on a minimal interesting value regarding the hypotheses and indicated sufficient power for 100 participants (50 men and 50 women). However, due to a mistake in the calculations the targeted power would have required 200 participants. Nevertheless, the power analysis revealed further insights about the distribution of answers and the test design. First, by removing the requirement that always half of the answer of each trial must be correct, more items can be analyzed. Second, especially for tests with only few trials there is likely a larger variance in the measured performance compared with the true ability resulting in lower standardized effect sizes for sex differences. These conclusions can also be derived from the calculations and insights of Brysbaert (2019) and Brysbaert and Stevens (2018), but some more details pertaining to the particular test and are provided in the following.

Expected mean accuracies for single item comparisons in trials with four mixed alternatives (VK tests) were set at .81 for men and .72 for women obtained by the estimation $p_2 = \sqrt[3]{s/n}$ from the results of Peters (2005) because of the large sample size and the same time limit of 3 minutes. As we assume comparable accuracies and reaction times for JJ tests and no sex differences, the estimate of .81 was used for both men and women for two paired alternatives. Note that these accuracy estimates already account for unanswered trials, that is, there is no distinction between wrongly answered and unanswered trials. The estimates for other combinations of trial designs could not be estimated from the literature and were based on minimal interesting effect sizes by assuming the smallest effect of the triple interaction to account for at least one sixth of the overall effect. To create a normal distribution of the data, the estimates were then transformed to logarithmic odds and the resulting values are presented in table 24.

Due to assumed large sex difference in VK tests (Cohen's $d > 0.7$), a standard deviation of 0.73 was assumed for all subtests, resulting in sex differences of $d = 0.7$ for four mixed alternatives. This standard deviation was further divided into a within-participants and a between-participants component based on the within-participant correlation of different mental rotation test scores of .66 reported by Voyer et al. (2006).

Based on these estimates, random normally distributed data was simulated in R (R Core Team, 2018) and the achieved statistical power was calculated for ANOVAs by the Superpower package (Caldwell & Lakens, 2019) and for linear mixed models by the lme4 package (Bates, Mächler, et al., 2015). For ANOVAs with 200 participants, the effect size of the triple interaction (sex*type of alternatives*number of alternatives) is estimated as $f = 0.14$ and the achieved power as .71. For the two-way interactions, the effect sizes are $f = 0.28$ and $f = 0.61$[15] and the estimated powers are >.99. With linear mixed models, the achieved power is comparable to ANOVAs. For 100 participants, the estimated powers for the two-way interactions are >.95, but the power for the three-way interaction drops to .41. The targeted power is only achieved for effect sizes of $f = 0.20$, which is, for example, achieved using the values in brackets in table 24.

[15] Note that under the assumption of no sex differences for two paired alternatives and large sex differences for four mixed alternatives, the sum of effects for two-way interactions has a lower bound but the effect size of the three-way interaction can be chosen freely. Disregarding the tests using eight alternatives, the sum of effect sizes for both two-way interactions is at least 0.6.

Table 24

Overview of Mean Estimate of Logarithmic Odds of Accuracy for Men (M) and Women (W).

Type/number of alternatives	2	4	8
mixed	M 1.45	M 1.45	M 1.45
	W 1.195	W 0.94	W 0.685
paired	M 1.45	M 1.45	M 1.45
	W 1.45	W 1.365 (1.44)	W 1.28 (1.43)

Note. Values in brackets represent values to achieve the targeted power for 100 participants.

For a more realistic power estimation, we further simulated binomially distributed results on only the rotated alternatives. As we intend to analyze individual items of each trial to account for the differences in the number of alternatives, the question arises, which distribution is appropriate. While results should be binomially distributed if the number of correct alternatives were random, participants know that always half of the alternatives are correct and the answers to individual items of one trial are not independent. Due to this symmetry of alternatives, the distribution on the correct and incorrect alternatives is the same and we analyze a binomial distribution on only the correct alternatives. Assuming the interdependence of answers to individual items is random, a binomial distribution for the answers to each item seems like the best approximation.[16]

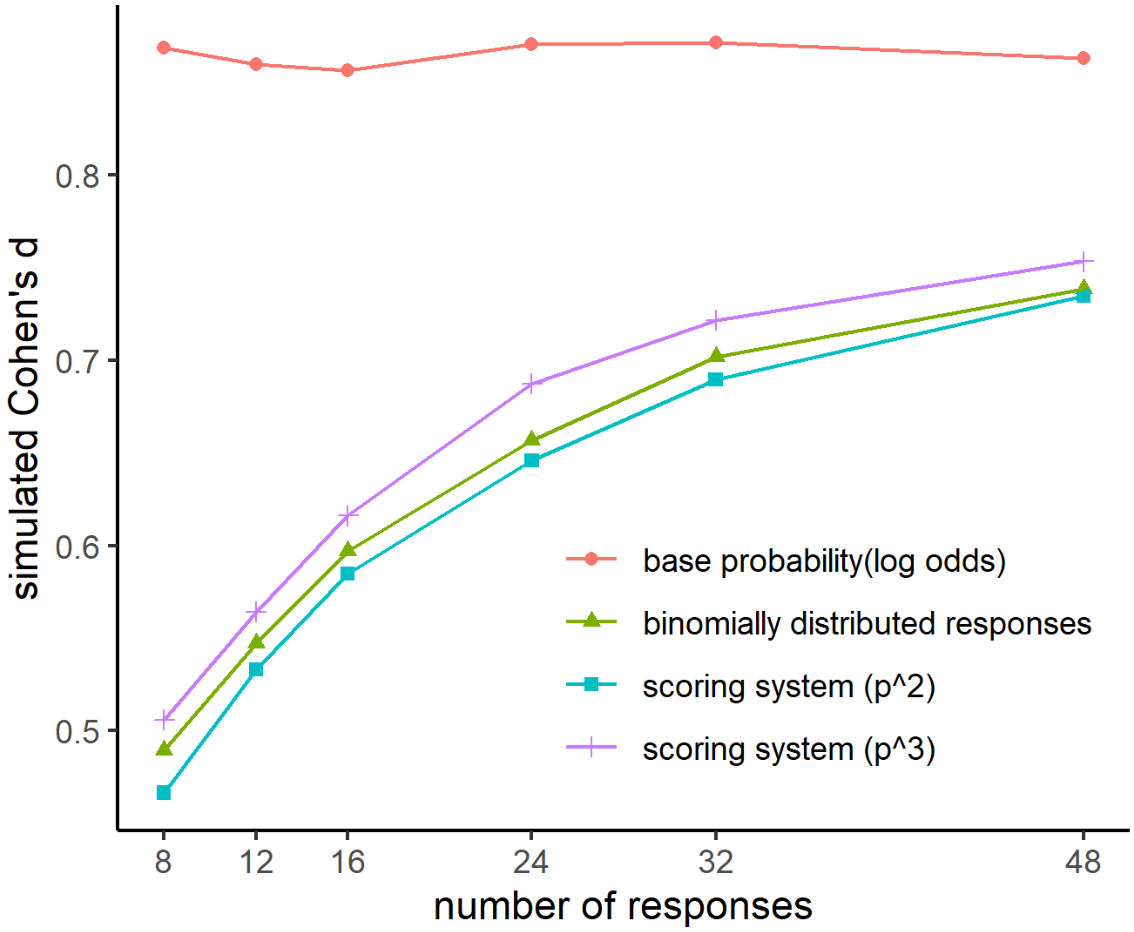
As the binomial distribution adds additional variance, we reduced the within-participants variance in the underlying logarithmic odds (but not the between-participants variance, as the correlation of .66 is obtained from actual observed data). Interestingly however, even when reducing the within-participants variance between trials to 0, the variance in the resulting binomial distribution of answers to 24 items is still larger than before, achieving on average Cohen's *d* of

[16] This may also by a disadvantage of the classical VK test. If the four alternatives were chosen at random, one could analyze four items per trial instead of only two.

0.66 for sex differences at four mixed alternatives. As can be seen in figure 31, Cohen's *d* increases with more responses as the relative variance of the binomial distribution decreases. The difference of 0.66 is still comparable or even lower than many observed large sex differences on 12 trials/24 items. However, as the variance of the binomial distribution depends on the probability there may be small variations depending on the scoring system. Assuming the probability to answer all alternatives on one trial correctly as either p^2 (based on independent answers) or p^3 (based on the strategies) produces very similar effect sizes as can be seen in figure 31.

Figure 31

Average Cohen's d of Sex Differences for 1000 Simulations and 100 Participants for Underlying Probabilities and Binomially Distributed Responses.



As it is unlikely that there is no variance within participants between different trials in the underlying odds, this implies fluctuations in the underlying probability between trials[17]. This indicates that controlling the difficulty to get better estimates of answer probabilities could increase power (but the extent cannot be estimated as there are no concrete estimates for the effect sizes).

A power analysis using generalized linear mixed models on the binomially distributed data then indicates comparable power to ANOVAs and linear mixed models if the number of responses is chosen such that the overall variances are similar. Due to the incorporation of covariates of item difficulty, a choice of test design parameters that have produced the largest sex differences in past research (cube figures, rotations in depth, mirrored distractors, time limit of 3 minutes), as well as general advantages in power of mixed models such as simultaneously accounting for by-participant and by-item variances, which could not be simulated, an acceptable power in the range of .8 should be achieved at N = 200 (or for the increased effect sizes as N = 100).

9.3.1.2 Participants

Despite online studies on mental rotation reporting comparable sex differences and standard deviations as laboratory experiments when accounting for the number of trials (J. K. Krüger & Suchan, 2016; Peters et al., 2007), there is the possibility of larger variance between participants in online studies due to inattention and less control of the testing environment. Accordingly, we increased the number of participants to 200 and included further between-subjects factors (experience with mental rotation tests, academic program), which have shown the largest between-subjects differences next to effects of sex (Peters, Laeng, et al., 1995), to reduce the variance within the sample. Moreover, we aimed for at least 50 men and 50 women without previous experience in mental rotation, as these were the originally intended target group and the overall most studied participants in mental rotation experiments. Participants were recruited

[17] For example, a binomial distribution of 100 items with probability .8 has a mean of 80 and a variance of 100*0.8*0.2 = 16, whereas a distribution of 60 trials with probability 1 and 40 trials with probability .5 also has a mean of 80 but a variance of 40*0.5*0.5 = 10.

through various newsletters at German universities and online platforms for psychological experiments. They received either study credit or could participate in a prize draw to win one out of four 25€ vouchers. Participation stopped after 48 men and 77 women without previous experience and 253 participants in total as this uneven distribution yielded the same power in the power analysis.

9.3.1.3 Measures

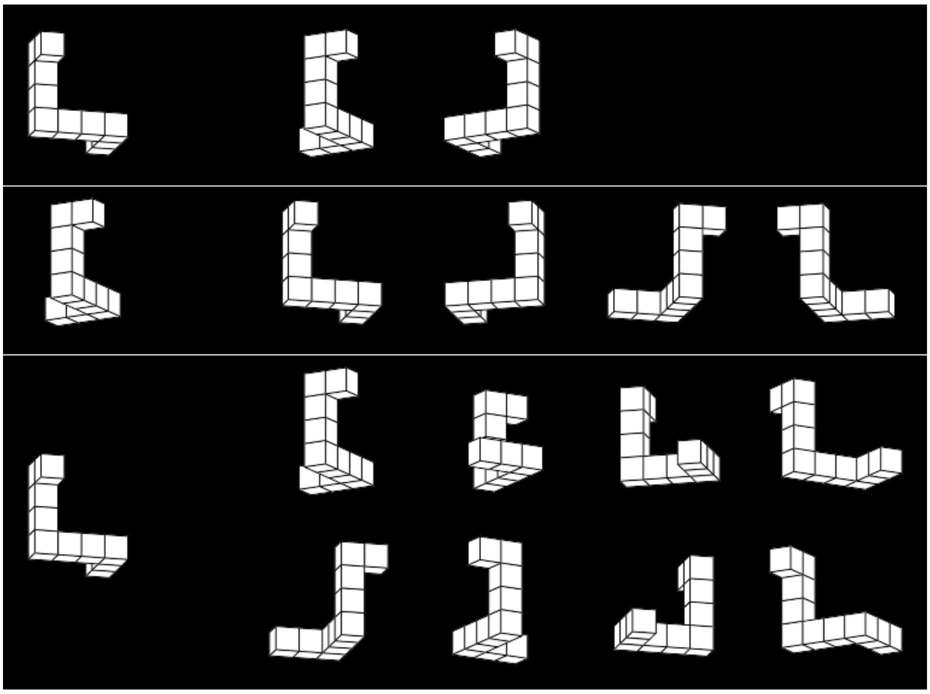
Participants were informed about the goals and procedure of the study before starting with the mental rotation test. The mental rotation test was succeeded by the demographic questionnaire. They were asked to plan sufficient time (30 minutes) and complete the test alone.

9.3.1.3.1 Mental Rotation

The mental rotation test was implemented using OSWeb (version 1.3.11) as part of OpenSesame (version 3.3.6; Mathôt et al., 2012) and made available online by JATOS (Lange et al., 2015). The test consisted of six blocks with a time limit of 3 minutes for each block. Each block consisted of trials of only one type of trials for a total of 48 alternatives (i.e., 24, 12, or six trials with two, four, or eight alternatives) in line with 12 trials for each block of VK tests. Trials were varied by the number of alternatives (two, four, or eight) and by the pairing of alternatives (paired or mixed). Every participant completed one block for each combination. In the paired condition, the alternatives were pairwise horizontally mirrored to one another and aligned with the canonical axes while the target was rotated compared with the canonical axes. In the mixed condition, the target was aligned with the canonical axes and the alternatives were ordered randomly. For all numbers of alternatives, the target was on the left side of the screen and the alternatives were ordered from left to right with an additional space of 50px (at a resolution of 1920x1080) between the target and the first alternative. In the case of eight alternatives, there were two rows of four alternatives each. Examples of trials are shown in figure 32.

Figure 32

Examples of Mental Rotation Trials With Two Mixed Alternatives (Top), Four Paired Alternatives (Middle), and Eight Mixed Alternatives (Bottom).



The order of the blocks was randomized for each participant. Within each block for each participant, the used cube models and rotation angles for the target were randomized such that each choice of parameters occurs again only after all other choices have occurred at least once. The orientation (of the target and alternatives) and angles of the alternatives were chosen randomly such that no two figures were the same and all alternatives differed from the target by angle. All randomizations were performed using inline JavaScript in OpenSesame using the modern version of the Fisher–Yates algorithm.

Table 25

Parameters for Stimuli Generation.

Parameter group	Parameter	Value
Color options	Background color	transparent (black)
	Border color	black
	Face color	white
Sizing and formatting	Cube Diameter	42px
	Image size	340px*340px
	File format	png
	Centering	optical
Model properties	Base orientations	a, b
	Models	2-8 and 12-16 (Peters & Battista, 2008)
	Base rotation angles (x, y, z)	-15°,0°,15° (a), -15°,0°,-15°(b)
	Angle difference	45°
	Rotational axis	z
	Order of rotation	z, x

The stimuli were generated using the library of the second study using the parameters given in table 25. Stimuli were rotated around the z-axis (vertical) first and the x-axis (horizontal) afterwards and the different orientations used different base angles such that the paired alternatives were actual mirror images of one another (see figure 32). For the structure of the cube figures, models 2-8 and 12-16 of the library of Peters and Battista (2008) were used as the other models can be transformed into these models by mirroring and/or rotating.

Participants could select and deselect the alternatives by clicking them with the mouse. Selected alternatives were marked by a quadratic white border. In the bottom right corner of the screen, there was a button to continue with the next trial. The button could only be clicked once exactly half of the alternatives were selected. Otherwise, the button text asked the participants to select exactly half of the alternatives and could not be clicked. There was no option to return to previous trials.

The test was preceded by three practice trials without a time limit in random order, one with two mixed alternatives, one with four paired alternatives, and one with eight mixed alternatives. For the practice trials, the participants received feedback in the form of a red or green border around the selected alternatives.

Before each block, participants were instructed about the number of trials, the number of alternatives, and the time limit, but not about the pairing of alternatives. Between blocks, they were allowed a self-paced break. As in VK tests, participants were instructed that they would get one point if and only if they selected all correct alternatives. During the trials, participants could see the number of the current trial and the number of overall trials for the block in the top left corner and the time left (since the start of the trial) in the top right corner of the screen.

9.3.1.3.2 Demographics

A digital questionnaire was used to collect demographic information. Participants were asked about their previous experience with mental rotation (participants had to indicate if they had or had not participated in other mental rotation experiments before), whether they were university students of STEM subjects specifically (no student at all, student in non-STEM subject, student in STEM subject), and their sex (male, female, or diverse). The resulting information is presented in table 26. All participants reported a cisgender identity but for two participants no answer was recorded, and they were removed from the dataset. Additionally, participants indicated their age in broad brackets in years (18-21: 33 men, 72 women; 22-25: 41 men, 68 women; 26-30: 14 men, 8 women; 31-35: 4 men, 6 women; or older than 35 years: 4 men, 1 woman), which was not part of statistical analyses.

Table 26

Number of Participants With Experience With Mental Rotation (Rows) and STEM Engagement (Columns), Grouped by Sex (M/W). The Additional Number of Outliers Are Given in Brackets.

Experience/STEM	No student	Non-STEM student	STEM student
no	M 3 (1)	M 29 (2)	M 13
	W 5	W 58 (6)	W 8
yes	M 0	M 40 (2)	M 5 (1)
	W 0	W 65 (5)	W 8

9.3.1.4 Statistical Analysis

The accuracy of each alternative was used as dependent variables and the number and pairing of alternatives, the sex, STEM education, and previous experience with mental rotation of participants and their interaction were used as independent variables. The number of alternatives was normalized to the interval [0,1] and treating the values as numerical was compared to treating the values as categorical. The angular disparity between the alternative and the target, the number of the block, and the position of the trial within the block were used as numerical covariates. The position within the block was normalized for each block to the interval [0,1] such that the last attempted trial had the value 1. In line with common scoring procedures, unattempted trials were treated as wrongly answered (with the missing values for angular disparity and the position within the block imputed by the respective mean value of that block). For trials, which were not finished due to the time limit, the answers were evaluated if at most half of the figures were selected. If more than half were selected, all answers to that trial were treated as unattempted.

Statistical analysis was performed with generalized linear mixed models with a binomial distribution using MixedModels package (version 4.0.0; Bates et al., 2021) in Julia (version 1.6.2; Bezanson et al., 2017). Model fit was calculated by using likelihood ratio tests to compare models with and without the effect of interest. Participants and cube models were used as random effects.

Random slopes were selected stepwise starting with a maximal model (to avoid over-parameterization at the start we included all two-way interactions for random slopes by participant and only the main effects for random slopes by the cube models) and removing random slopes by dropping variance components using an LRT backwards heuristic at $\alpha = 0.2$ (Matuschek et al., 2017). Non-significant fixed effects were further stepwise removed from the model, such that effects that least decreased model fit were removed first and a model containing only significant fixed effects remained. Non-significant effects were then tested for an improvement of model fit by inclusion in the resulting model, while significant effects were tested for worsening of model fit by exclusion of the effect. The resulting p-values were compared to a significance level of .05. The analysis of main effects contained in significant interactions was performed according to Levy (2014). The numerical variables were centered around the mean and normalized to range 1 and the categorical variables used sum contrasts. Due to the skewed distribution of STEM education, the sum contrasts for STEM education were further centered around their respective mean because this produces estimates for the average effect over all participants instead of the grand average of the group means.

While there are several advantages to mixed models, the internal optimization procedures are not exact, which produces large imprecisions in both estimated effect sizes and p-values when the random effects are overparametrized. In addition to the procedures of Matuschek et al. (2017), we have reduced the random effects structures until these uncertainties were of magnitudes of 0.001 for the p-values and point estimates in our checked samples.

To supplement the analysis, Bayes factors were calculated using the approximation of Wagenmakers (2007) and compared to the decision boundary factor 3 or $\frac{1}{3}$. They should, however, be treated with caution as there is a monotonous relationship between these Bayes factors and p-values (see the fourth study). For both frequentist and Bayesian analyses there is ongoing discussion about the optimal procedure and we release all data and code in accordance with the suggestion of Matuschek et al. (2017).

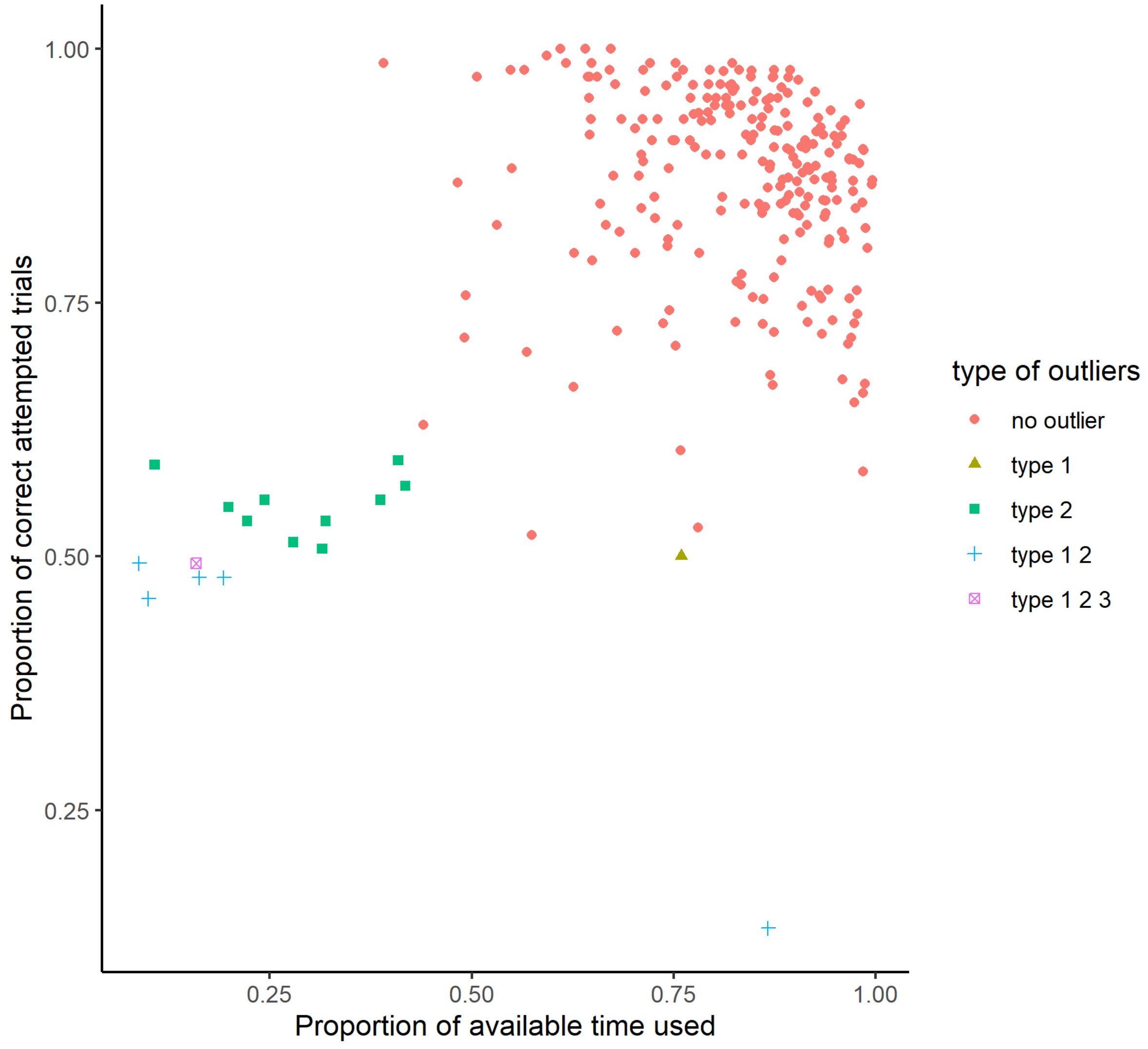
9.3.2 Results

9.3.2.1 Outliers

Four procedures were implemented to detect outliers. Overall, 17 participants (see table 26 for the demographics) were deemed outliers and excluded from analysis. Seven participants were excluded because their overall performance was below chance level (0.5) on their attempted trials. Sixteen participants (six of them already excluded by the first procedure) were excluded because the sum of their relative time used and their accuracy on their attempted trials was below 1, which indicated a too strong focus on finishing the test quickly instead of accurately. One participant selected the first (right) answer in more than 90% of all trials indicating answering in a pattern instead of correctly but that participant was already excluded by both previous procedures. No participant attempted less than half of all trials. The resulting speed-accuracy trade-offs of participants and outlier detection is shown in figure 33. These indicate that many participants were limited by either time or accuracy (ceiling effects). But there are some further participants towards the lower left corner of figure 33, who were likely neither limited by time nor accuracy and could have performed better. While the consideration of motivation as a limiting factor of performance and further research into exclusion criteria based on performance is interesting, we refrained from excluding further participants as our outlier detection was already more restrictive than most outlier detection measures implemented in other studies.

Figure 33

Speed-Accuracy Trade-offs and Outliers.



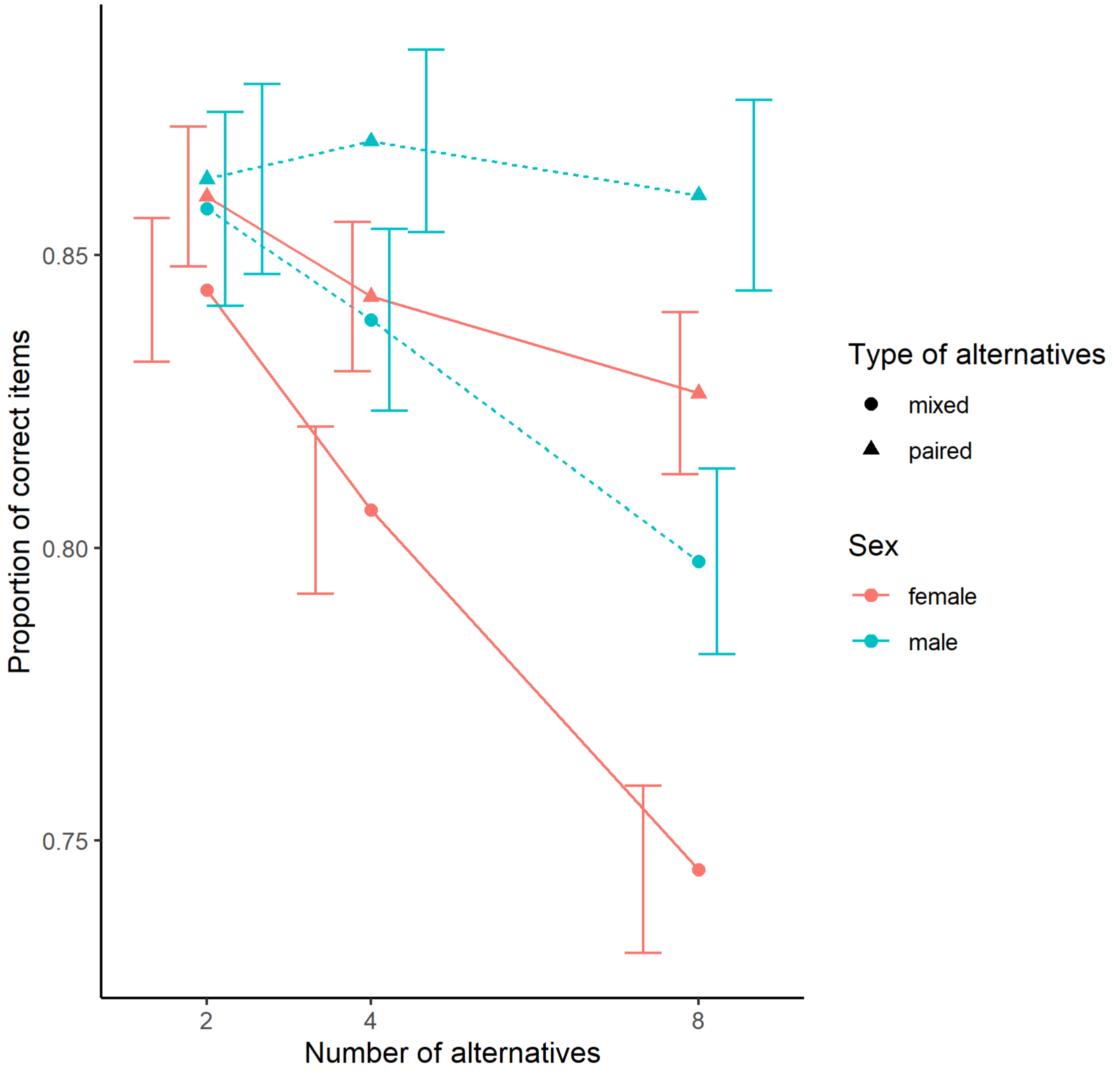
Note. Type 1 outliers – performance below chance level. Type 2 outliers – Too strong focus on speed over accuracy. Type 3 outliers – almost always selecting the first alternative.

9.3.2.2 Scores

The descriptive data of the proportion of correctly solved mental rotation items is shown in figure 34 for all participants and for the subgroups of experience and STEM education in figure 35. Descriptively, the data is in line with the main hypothesis: Women perform worse than men for the mixed alternatives, for more alternatives, and for the combination of both.

Figure 34

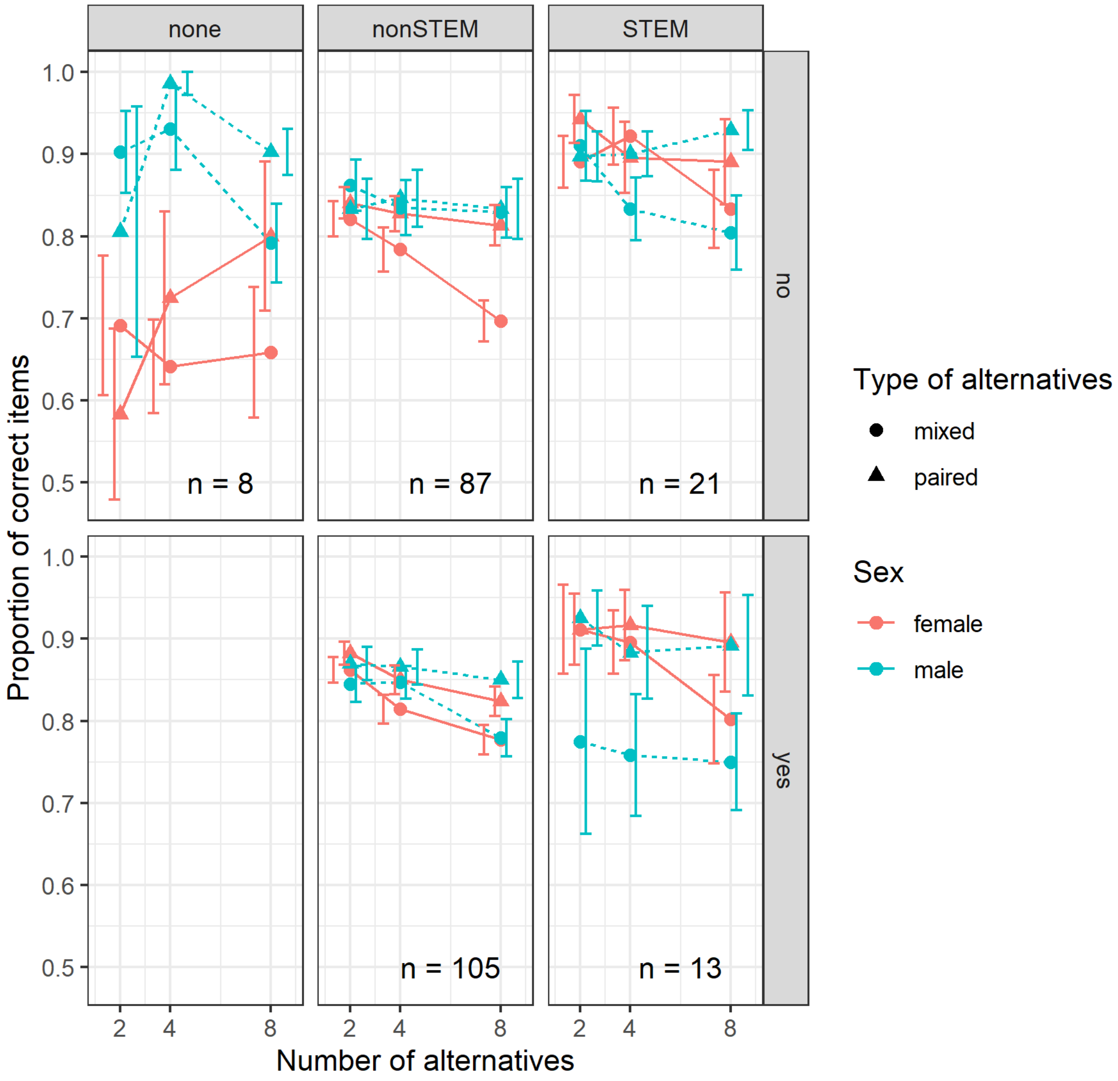
Proportion of Correctly Solved Mental Rotation Items (Mean and Standard Error), Separated by Number and Type of Alternatives.



The results do, however, vary strongly for the different combinations of previous experience with mental rotation and STEM education. The large fluctuations for the STEM students and non-students may in part be due to the low number of participants for those demographics.

Figure 35

Proportion of Correctly Solved Mental Rotation Items (Mean and Standard Error), Separated by Number and Type of Alternatives, and the Participants' Education (Left/Right) and Experience (Top/Bottom).



For the mixed model, the model selection resulted in a model including random slopes for the number of alternatives (as categorical), the type of alternatives, the number of the block, angular disparity, and position of the trial within the block, and random intercepts by participant and random slopes for experience and random intercepts by cube models. The random slope for angular disparity by participants was further removed because the resulting model showed larger than intended variations and this least decreased model fit. As fixed effects, the number of alternatives was treated as numeric because the model using categorical data was not better ($\chi^2(20)=24.21$, $p=.233$).

Overall, the five-way interaction was significant, but the partial interactions of interest regarding sex differences by the trial design were not significant. For the effects of STEM education, experience, and their interaction with sex, only the main effect of STEM education was significant. We further analyzed the interactions of the main hypothesis separately for all participants without previous experience and the non-STEM students specifically, because these were the participants of main interest and the other combinations of STEM education had too few participants.

For both the participants without previous experience with mental rotation overall and the subset of non-STEM students in particular, there were significant sex differences, which interacted with the type of alternatives. Men performed better than women overall and this effect was larger and significant for the mixed alternatives compared with the paired alternatives, for which the sex differences were not significant.

In all analyses, participants performed better on tasks with fewer alternatives and this effect was larger for the mixed alternatives than for the paired alternatives. Participants improved between blocks, but decreased performance was found within blocks for later trials. Items with larger angular disparity were solved correctly less often.

The Bayes factors were inconclusive regarding most of the significant effects regarding sex differences and require more evidence.

Table 27

Statistical Analysis of (Logarithmic Odds of) the Proportion of Correct Answers.

Variable	Estimate	SE	Test statistic	p	BF
intercept	2.16	0.13			
STEM*exp*nAlternatives*type*sex			$\chi^2(1)=8.51$	.004	0.22
nAlternatives*type*sex	0.04	0.18	$\chi^2(1)=0.21$	.645	13.77
nAlternatives*type	-0.48	0.09	$\chi^2(1)=26.40$	<.001	<.01
nAlternatives*sex	0.15	0.12	$\chi^2(1)=1.26$	.261	8.15
type*sex	0.07	0.11	$\chi^2(1)=0.48$	.487	12.03
sex (male-female)	0.21	0.14	$\chi^2(1)=3.75$	.053	2.35
nAlternatives	-0.45	0.06	$\chi^2(1)=41.62$	<.001	<.01
type (mixed-paired)	-0.34	0.05	$\chi^2(1)=33.98$	<.001	<.01
STEM*exp*sex			$\chi^2(1)=0.20$	.655	13.84
STEM*sex			$\chi^2(2)=4.66$	.098	22.77
exp*sex			$\chi^2(1)=2.11$	.147	5.33
STEM*exp			$\chi^2(1)=1.28$	.258	8.07
STEM			$\chi^2(2)=11.76$	.003	0.65
Exp(yes-no)	0.14	0.16	$\chi^2(1)=2.37$	.124	4.68
block	0.98	0.09	$\chi^2(1)=98.87$	<.001	<.01
deg	-0.60	0.05	$\chi^2(1)=147.22$	<.001	<.01
trial	-0.92	0.07	$\chi^2(1)=118.84$	<.001	<.01
Participants w/o experience					
STEM*nAlternatives*type*sex			$\chi^2(2)=12.68$	.002	0.20
nAlternatives*type*sex	-0.04	0.26	$\chi^2(1)=0.01$	.908	10.72
nAlternatives*type	-0.63	0.13	$\chi^2(1)=21.54$	<.001	<.01
nAlternatives*sex	0.08	0.17	$\chi^2(1)=0.19$	.664	9.79
type*sex	0.29	0.15	$\chi^2(1)=3.85$	.050	1.57
type(mixed)*sex	0.63	0.22	$\chi^2(1)=8.16$	.004	0.18
type(paired)*sex	0.25	0.22	$\chi^2(1)=1.27$	.259	5.71
sex (male-female)	0.42	0.20	$\chi^2(1)=4.28$	.039	1.27
nAlternatives	-0.36	0.09	$\chi^2(1)=12.81$	<.001	0.02
type (mixed-paired)	-0.26	0.18	$\chi^2(1)=9.07$	.003	0.12
Non-STEM students w/o experience					

Variable	Estimate	SE	Test statistic	p	BF
nAlternatives*type*sex	0.39	0.30	$\chi^2(1)=1.63$	.201	4.13
nAlternatives*type	-0.74	0.14	$\chi^2(1)=27.47$	<.001	<.01
nAlternatives*sex	0.13	0.21	$\chi^2(1)=0.37$	.542	7.75
type*sex	0.39	0.17	$\chi^2(1)=4.62$	.032	0.93
type(mixed)*sex	0.80	0.26	$\chi^2(1)=8.05$	.005	0.17
type(paired)*sex	0.30	0.25	$\chi^2(1)=1.34$	.248	4.77
sex (male-female)	0.48	0.23	$\chi^2(1)=4.03$	.045	1.24
nAlternatives	-0.43	0.10	$\chi^2(1)=14.64$	<.001	<.01
type (mixed-paired)	-0.21	0.09	$\chi^2(1)=4.26$	.039	1.11

Note. Exp – previous experience with mental rotation, nAlternatives – number of alternatives, type – type of alternatives, block – number of block, deg – angular disparity, trial – position of trial within block, BF – approximated Bayes factor in favor of the null hypothesis.

9.3.3 Discussion

This experiment provides some evidence that sex differences in mental rotation tests are at least partially due to the trial layout, especially due to the mixed presentation of alternatives instead of a pairwise mirroring. The sex differences due to the trial layout interacted with both STEM education and previous experience. In line with the preregistered recruitment, the group of participants of main interest were those without previous experience with mental rotation. These, and especially the subgroup of non-STEM students, are the most tested sample in other studies, which provided the evidence for large sex differences. For these participants, the results show that sex differences are larger for mixed alternatives and smaller and not significant for paired alternatives in line with the hypothesis and the small or non-differences found for SM and JJ tests.

Descriptively, the number of alternatives also interacted with sex differences and these results were also observed in the overall sample but were not significant. Due to the low power, these effects need to be further investigated. In the overall sample, a possible reason is the broad measure of experience. Contrary to the hypothesis, experienced participants did not perform significantly better, but there could be a large variance in specific experience between participants.

Both differences in magnitude and in experience with specific tests could transfer differently to performance for specific trial layouts. In line with the hypothesis, education had a significant effect on performance, but the number of STEM students and non-students was too low to draw meaningful conclusions.

Effects of the trial layout were also observed independent of sex. Participants performed worse in trials with more alternatives and for mixed alternatives and especially for the combination of both. However, these effects were smaller than the effects of the other covariates. In line with the hypotheses, items with smaller angular disparities and trials in later blocks were solved correctly more often. Two conclusions can be drawn from this. First, the improvement between blocks and the large number of blocks compared with traditional VK tests lead to a higher-than-expected overall accuracy and could have reduced the chances for the detection of effects due to ceiling effects. Second, in line with the prediction in the first part that overall accuracies are quite comparable between SM and VK tests, the variations between trial layouts are negligible compared with the variations within the trials due to angular disparity. However, the overall effect of trial layout on task difficulty should be further explored. Another point for discussion is the decreased performance within blocks contrary to the hypothesis, despite the exclusion of unattempted trials at the end of blocks. One explanation could be varying speed-accuracy trade-offs with a larger focus on speed in later trials due to the approaching time limit.

9.4 General Discussion

The theoretical and experimental investigation of sex differences in mental rotation test performance provide promising evidence that they are due to the trial layout within tests. In the first part, we have reviewed research on influence factors on sex differences in VK tests. However, as many of these factors failed to influence sex differences in the same way in SM or JJ tests, we have identified the trial layout as a promising possible reason for sex differences. By comparing the trial layouts of existing tests, the differences rely on mostly two factors: the number of alternatives and whether the alternatives are presented as pairwise mirrored or mixed. Especially the type of alternatives has been shown to indeed influence sex differences in the experimental investigation.

9.4.1 Sex Differences in Mental Rotation Test Performance

Our results suggest that the large sex differences in VK tests are not due to the abilities measured by SM or JJ tests but due to additional abilities involved only in solving the trials of VK tests. Much research on mental rotation performance has focused on different strategies for comparing two figures with varying evidence for sex differences in the use of strategies (Hegarty, 2018; Khooshabeh & Hegarty, 2010; Scheer et al., 2018; Voyer et al., 2020). In fact, Voyer et al. (2020) found no evidence for existing theories on strategic differences between sexes and suggested "that research needs to look elsewhere to account for these sex differences" (p. 887). Our results suggest such a look elsewhere. Whereas strategies of comparing two rotated figures might be useful to analyze individual performance differences in all mental rotation tests, the reason for sex differences could instead lie in the strategies used to disentangle mixed alternatives. Similarly to strategic differences, effects such as stereotypes have been suspected to influence sex differences in mental rotation performance. Again, this influence seems to only affect the multiple-choice strategy of disentangling mixed alternatives instead of the mental rotation ability measured by SM or JJ tests.

Mental rotation ability is defined as the ability to rotate objects in the mind, and it is a reasonable argument that this ability is measured by SM or JJ tests. The additional abilities involved in solving mixed alternatives in VK tests should thus be the reason for the larger sex differences. However, it is not clear what they are. One possible link is the involved visual working memory. Hyun and Luck (2007) identified the object working memory subsystem and not the spatial working memory subsystem to interfere with SM test performance. Kaufman (2007), however, identified the spatial working memory subsystem to fully mediate the relationship between sex and spatial ability in general and partially for VK test performance. The additional involvement of spatial working memory compared with SM tests could therefore be one reason for sex differences in VK tests. Indeed, a male advantage in visual-spatial working memory is supported by meta-analytic results (Voyer, Voyer, et al., 2017). A possible link could thus be that the spatial layout of the increasing number and complexity of alternatives of VK trials taps into the spatial working memory. There are, however, still open questions regarding the involvement of visual working memory. First, Kaufman already addressed the need for further research for the larger and unexplained effects for VK tests compared with spatial ability in general. Second, the results of Hyun and Luck (2007) were only obtained for the rotation of a single letter and should be replicated with more common mental rotation tests. Moreover, the involvement of spatial working memory through the trial layout cannot explain the interestingly large sex differences for SM tests using two-dimensional polygons (Heil & Jansen-Osmann, 2008; Jansen-Osmann & Heil, 2007b). It is unclear, whether the complexity of the polygons also taps into spatial working memory or if there are also other mechanisms involved. Furthermore, spatial working memory also cannot explain the interaction of sex differences with many other factors reviewed in the first part such as gender stereotypes.

Despite them not being clearly related to the mental rotation process, this does not mean that work on previously identified mitigating factors on sex differences in psychometric tests are in vain. Further investigating sex differences in psychometric test performance is nevertheless interesting as these are some of the largest in cognitive psychology. Our understanding of these sex differences can enhance our understanding of cognitive sex differences in general.

9.4.2 Mental Rotation Trial Design

The review of trial designs also provides some insights as well as possible directions for further research. The influence of the type of alternatives further enhances the conclusions of the second study that the abilities involved in the JJ test are the same as those in the SM test. In contrast, presenting two arbitrarily mirrored alternatives could have involved the same additional abilities of VK tests. The geometric discussion on mirroring planes and trial layout of the second study thus also seems to be of importance. The possibility to identify alternatives as mirrored to one another could also be influenced by the choice of the stimulus material. Two mirrored figures can always be transformed into each other by a mirroring plane, but this mirroring plane might be arbitrarily skewed in three-dimensional space. For both more realistic stimuli and real three-dimensional objects it could be easier to identify them as mirrored to one another, as they promote a three-dimensional viewpoint or real three-dimensional experience with them might make it easier to deal with skewed mirroring planes. This in turn could explain the possibly reduced sex differences for these stimuli.

As was done in the experiment, systematically parametrizing different trial layouts also allows the isolation of parameters of interest with sex differences. The theoretical review also allows the incorporation of further parameters. These can not only be explored for their link to sex differences but could also help to identify and separate other involved abilities, isolate what we understand as mental rotation ability, and additionally increase the power of experimental designs.

9.5 Limitations

The theoretical review of sex differences is limited by the facts that there was no systematic identification of varied factors between test designs and there was no systematic literature search or meta-analytic comparison for each assumed factor. This was in part due the large and non-systematic variation of test designs between studies, where studies often differ in multiple design parameter even when employing similar tests. As has been outlined, the varied parameters were often also not the target of the investigation. Between SM and VK tests, the variations are much larger and the different scoring systems further hinder a direct comparison. Another problem might be a publication bias in the reviewed literature favoring the publication of larger sex differences and only significant explaining factors. While this could lead to an overestimation of sex differences in VK tests, a possible overestimation of sex differences in SM tests or neglecting factors which failed to reduce sex differences in VK tests does not impact the conclusions drawn here.

The main limitation of the experiment is that too few participants were tested for the desired power. The significant results thus need to be replicated and the non-significant results need to be further investigated with more participants. Between subjects, the factors STEM education and experience need further investigation. While the identified design factors can explain sex differences for the most studied population of social/human science students without mental rotation experience, no clear conclusions can be drawn whether these mechanisms hold for the general population.

There are also two factors identified in theory, which could not be investigated in the experiment. Here, we only investigated multiple comparisons within one trial through more alternatives within one trial, but this cannot be separated from effects of more possible pairwise comparisons through more alternatives in general through the presentation of multiple trials at once. Another possible limitation is linked to the strategies of solving alternatives in VK trials. One assumed strategy was that participants always solve the last alternative by exclusion. By using such a strategy for all trials, the overall number of necessary comparisons increases with the numbers of alternatives. Similarly, pairing the alternatives could have reduced the number of necessary comparisons. The identified effects could thus be attributed to effects of a relatively stricter time limit.

9.6 Conclusion

In conclusion, the present study provides evidence that the trial layouts of different mental rotation tests are the reason for varying sex differences in respective test performances. Especially presenting alternatives as pairwise mirrored could mitigate the male advantage. The theoretical implications of different trial layouts could help to further understand mental rotation ability as well as other cognitive abilities involved in solving the tests.

10 General Discussion

10.1 Summary

Over the course of six studies, we have attempted to gather insights about various aspects of mental rotation. The first and the fourth study focused on the link to physical activity. The first study focused on the interactions of simultaneous mental rotation and aerobic exercise in the form of cycling. In the fourth study, we further explored the link between mental and physical rotation by isolating the visual rotation occurring within all manual rotation trainings. To improve the interpretation of mental rotation test results, the second and the sixth study were mostly concerned with the design of mental rotation tests and the third study was concerned with the analysis of results. In the second study, we evaluated the analysis of all mental rotation trials by incorporating a second alternative compared with the analysis of only half of all trials in traditional chronometric tests. As we also found improvements during the mental rotation session, we explored the potential advantages of including practice effects in the analysis. The sixth study combined a comparison of different test designs with an investigation of sex differences in performance. Sex differences were also analyzed in the fifth study in relation to the implicit evaluation of the stimulus material.

The results of the first study indicated that both physical and cognitive performance could be maintained but at the cost of a higher subjective cognitive effort and a higher physiological effort. Regarding the specific relationship between mental and manual rotation, the results of the fourth study indicated that the visualization of rotation could be the main link. We separated the visual rotation and the congruent and causal rotational hand movement from manual rotation trainings. Neither the causal nor the congruent rotational manual actions caused additional training effects.

In line with our predictions for the second study, the additional alternative produced the monotonic and almost linear relationship between the angular disparity and the reaction time for all stimuli, which had previously only been established for the rotated trials. The further comparison of mental rotation tests in the sixth study revealed a useful parametrization of similarities and differences between test designs. Next to the test design, we focused on the data analysis in the third study. The dataset of the second study was used again with a focus on practice effects within sessions and their impact on the interpretation of results in group comparisons. The findings suggest an improved detection of treatment effects by separating them from the practice effects within sessions.

Regarding sex differences in mental rotation performance, the fifth study revealed the implicit evaluation of the commonly used cube figures as one possible reason. This result did, however, not appear with pellet figures. The further investigation of the test design and their relation to sex differences in the sixth study indicated an effect of the layout of individual trials especially by the pairing of alternatives. In the following, the main effects will be discussed in more detail.

10.2 Physical Activity and the Training of Mental Rotation

Due to the link between spatial abilities and STEM performance, the enhancement of mental rotation performance is an interesting topic. Mental rotation has been shown to improve both after repeated mental rotation practice (e.g. Heil et al., 1998; Meneghetti et al., 2017; Peters, 2005; Peters, Laeng, et al., 1995) as well as manual rotation practice (Adams et al., 2014; Wiedenbauer et al., 2007; Wiedenbauer & Jansen-Osmann, 2008). Whereas Heil et al. (1998), Meneghetti et al. (2017), and Peters, Laeng et al. (1995) found improvements over multiple sessions of psychometric and chronometric tests, Peters (2005) also demonstrated better performance in the second block of a single psychometric test. The results of Adams et al. (2014) further indicate that manual rotation produces similar training effects as mental rotation practice. In line with these results, the first, second, and fourth study also demonstrated an improved performance during single sessions of chronometric tests and the sixth study also for an improvement between multiple blocks of one test within one session. In the fourth study, we also replicated the performance enhancements of manual rotation training. However, these improvements did not differ between the congruent manual rotation, the causal manual rotation, and the only visual rotation. Thus, it seems plausible that the visual rotation whether internal (imagined) or external (physical or visualized) is the main reason for improvements. Especially the visual rotation thus allows further and easier training methods to enhance mental rotation performance because it does not require manual rotation devices and is more adaptable as the onset and speed of the visual rotation are independent of the manual rotation. On the other hand, the separation of the phases of the manual rotation can be used to gain more insights about the processing stages of mental rotation.

The results can further be combined with the general link of simultaneous cognitive and physical activity explored in the first study, where both cognitive and aerobic performance could be maintained. Despite the unclear link between training load and training effect for both mental rotation and aerobic exercise, the simultaneous training of both could prove time efficient. Adams et al. (2014) already demonstrated that the manual rotation training can benefit both mental and manual rotation performance with no disadvantage for the mental rotation performance compared with the isolated mental training. Thus, a combined physical and cognitive training could prove fruitful compared with isolated practice. However, the increased effort, likely due to neural resource conflicts during simultaneous exercise (Audiffren & André, 2015; Baumeister et al., 1998, 2007; Dietrich & Audiffren, 2011; Englert, 2016), could impact motivation and the feasibility of such a training.

10.3 Mental Rotation Tests

Traditional mental rotation tests rely on the discrimination between rotated and mirrored stimuli. As the usefulness of mentally rotating mirrored stimuli is questionable, they are typically excluded from analyses of chronometric tests. This means that additional time is necessary to collect the answers discarded in the future or a loss of power due to fewer analyzable responses (Brysbaert, 2019; Brysbaert & Stevens, 2018). To make mental rotation the most useful strategy in every trial in the test of the second study, we used two alternatives, which are mirrored to each other. The target was thus always rotated to one alternative. The results indeed demonstrated the monotonic and almost linear relationship between angular disparity and reaction time, which is generally seen as an indication of mental rotation. These results were again verified in the fourth study. In both studies, however, there were some differences in both reaction time and accuracy when the answers were separated by targets as well as by the rotation axes indicating some possibilities and the need for further research.

Aspects of the design were again analyzed in the sixth study by comparing the design of the second study to other tests. This comparison verified the usefulness of incorporating more analyzable trials to increase the statistical power. Moreover, it revealed some theoretical possibilities to further improve the trial layout. These could be especially useful for psychometric tests, where additional aspects of strategies within trials come into play. The further possibilities to systematically vary the trial layout could help to isolate aspects of the tested abilities.

The empirical results of the sixth study verified an influence of the trial layout on overall performance at least for the unexperienced participants. A possible reason could be the use of strategies within trials as well as the combination of mental rotation with other cognitive abilities. Most notably, the compared trial layouts confirmed that the choice of pairwise mirrored alternatives based on theory in the second study was indeed of practical relevance and a sensible choice. However, it must be further studied whether and how much mental rotation is used to solve the trials. Whereas the investigation of manual rotation, as in the fourth study, can help our understanding of the mental rotation process, the combination with questions of the test design is equally important as this has now also been demonstrated to influence performance.

Next to improvements of the test design itself, the results of the third study suggest that an incorporation of practice effects can prove useful in isolating effects of an intervention in pre-post designs. Next to interesting aspects about practice effects themselves, the inclusion in statistical analyses can help separate them from learning effects due to an intervention. The random simulations suggest that such a mix of practice effects and intervention effects can be an actual obstacle and not only of theoretical concern. The incorporation of practice effects in the analyses can thus increase the power to detect the actual effects of interest.

10.4 Sex Differences in Mental Rotation Performance

Sex and/or gender differences in psychometric test performance are some of the largest known in cognitive psychology. The interest in this topic is further enhanced due to the link to STEM performance and the lower female engagement in STEM topics. The most widely used cube figures are seen as more male stereotyped and have been suspected as one source of performance differences (Rahe et al., 2020; Rahe & Quaiser-Pohl, 2020; Ruthsatz et al., 2014, 2015). While this may be true on a population level, the results of the fifth study revealed that on the individual level the implicit affective evaluation of the cube figures instead of the explicit gender stereotype of the figures was the influential effect. However, as we did not detect significant differences in neither mental rotation performance nor in the implicit evaluation, it is not clear whether the more male stereotyped nature of the cube figures also causes sex differences in the implicit evaluation on a population level and thus mental rotation performance differences. Moreover, while the implicit evaluation of the cube figures was associated with the test performance and offers a promising link between performance and emotion, creating more neutral stimuli may not be able to reduce sex differences as the relationship to the implicit evaluation was not observed for the pellet figures.

As sex differences are negligible or much smaller in chronometric tests and were also not observed in the fourth study for the trial design of the second study, the layout of the individual trials may be a reason for sex differences in performance. This was investigated in the sixth study and the results did indicate such an effect. Especially presenting them pairwise as mirrored could reduce performance differences between sexes. This suggests that the sex differences do not occur in the mental rotation ability itself but in other abilities involved only in the solution of psychometric tests, which need to be identified further. Nevertheless, it remains interesting why some overarching aspects of psychometric tests, such as the relationship between the implicit attitudes and the stimulus material identified in the fifth study, are related to test performance and what measures can be implemented to reduce sex differences.

10.5 Limitations

Next to the individual limitations of each study, there are some overall limitations of this thesis. First, all studies used participants of mostly similar demographic characteristics (regarding, e.g., the university, study program, mother tongue, and age) and some participants may have participated in more than one of these studies. While the second problem was partially circumvented in the sixth study by focusing on participants with no previous experience with mental rotation, the transfer to the general population is not entirely clear. Another limitation arises from the fact that all participants were cis gender. Despite the interest in sex and/or gender differences in performance, we could not distinguish between gender and sex as the possible reason in our studies.

This thesis is also limited by the fact that the studies dealt with multiple topics instead of a single topic in depth. While there are some connections and shared insights about mental rotation, especially the first and fourth study could have profited from follow-up studies adding more depth to the research. Unfortunately, some of the planned studies had to be postponed due to the COVID-19 pandemic. Despite the desire to analyze one aspect in its entirety, this thesis demonstrates the variety and interconnectedness of research regarding only one specific cognitive ability and how the combination of multiple aspects is necessary to facilitate our understanding of this ability. This problem is also relevant on larger scales: For example, as mental rotation is embedded within spatial abilities, a further understanding of all spatial abilities could be crucial to further understanding specific abilities. And as has been reviewed, there is still much open discussion even regarding only the classification of and thus also the relationships and transferability between spatial abilities. As such, this thesis highlights our current limited knowledge and the complexity of both mental rotation and spatial abilities, which require much further research.

Moreover, the quite recent discussion on the replication crisis in psychology has highlighted some methodological advancements and practices to generate reliable results. However, as I only learned about the implications of some of these methods throughout this thesis, they could not be applied in all studies. Especially the first and second study could have profited from more participants and preregistered analyses. This problem is reduced for the second study as the results were replicated in the fourth study. Furthermore, as I am also not free from biases and suffer from constraints due to the framework of the scientific system, the discussions may lack some depth to adequately acknowledge the uncertainty and possible contradicting conclusions. All implications should be treated with caution and in light of possible future advancements and insights about the currently used methods.

10.6 Outlook

The presented studies offer much potential for further research both by extending the results as well as further linking them together. Whereas cognitive performance is mostly assumed to be limited by speed, accuracy, and trade-offs between them, the results of the first study indicate that cognitive effort might play a role. There is the possibility for further measurements through pupillometry to verify such effects. Pupil diameter during cognitive tests has been shown to be a measure of cognitive activity (Beatty & Lucero-Wagoner, 2000). This if often labelled as cognitive effort but the distinction from cognitive load or demand is not clear. Moreover, pupillometry can only identify changes relative to other states such as a baseline state, but there is no distinct value for maximal or minimal effort. At least for chronometric tests, we have found no evidence of different cognitive effort between men and women (Bauer, Jost, & Jansen, 2021; Bauer, Jost, Günther, et al., 2021). This further enhances the conclusions of the sixth study that the large sex differences in psychometric mental rotation tests do not emerge from the abilities involved in the chronometric tests. Regarding the test designs, it can be further investigated whether different trial layouts require different cognitive effort from men and women as a possible reason for sex differences.

Sex differences themselves require further attention and it is even unclear if they should rather be described as gender differences. As in part outlined in the fifth and sixth study, there are a multitude of suspected reasons including social explanations, biological differences, as well as cognitive differences in strategies and spatial working memory. For most of these, it is unclear why or how they interact with the reduction of sex differences due to aspects of the test design, such as time limits or the type of distractors, which have been found to affect sex differences. The implicit affective evaluation of the stimulus material investigated in the fifth study provided one possible underlying explanation for effects of the stimulus material. But the affective evaluation and its influence on mental rotation test performance itself also needs further research. Whereas we have used different stimuli and correlated the affective evaluation with performance, it would be interesting whether and how the affective evaluation of one stimulus can be manipulated and whether such a manipulation would cause the same effects. The sixth study also introduced the trial design, for which it can only be hypothesized why it affects sex differences and how it interacts with the suspected explanations. Manipulations to increase sex differences, such as polygons as stimuli in psychometric tests, could also be employed to enhance our understanding of them. Similarly, a manipulation of affective evaluations could also be investigated to increase sex differences in mental rotation test performance.

Another possibility to reduce sex differences is through training, but overall training effects should also be investigated further. Next to improvements in speed and accuracy, an interesting question is whether cognitive effort is also reduced. However, as participants rarely reported maximal effort but still had room for improvement, another improvement through training could be to enable people to invest more cognitive effort into one task. Regarding the improvements of performance due to treatments, the analysis of learning within sessions of the third study also allows further interpretation. For example, an increased learning speed in the posttest could indicate a useful interaction of the treatment and the repetition of the task. On the other hand, a reduced learning speed in the posttest could indicate ceiling effects as a possible reason for null effects of treatments. Separating the phases of mental rotation by the analysis of manual rotation could further help to pinpoint where improvements occur. Of course, the studies open many possibilities of training methods through variations of the training parameters but also through the combination with aerobic exercise. Next to the magnitude of learning effects, it should be checked whether improvements also transfer to other spatial abilities and STEM performance, especially as a means to reduce the gender disparity in STEM fields.

The studies also provide many possibilities to vary the trial layout as well as the stimulus material of mental rotation tests. Further studies could also explore influences on mental rotation performance of different cube figures or different versions of the same figures. This includes the expansion in the three dimensions, the alignment with the origin, and coloring parameters. Design parameters could also be varied in attempt to force different strategies and eye tracking or EEG could be used to analyze the actual use and possible sex differences in the choice of strategies. The link to manual rotation should also be explored further. Applying the analysis of the different phases of manual rotation of the fourth study could help to differentiate between the stages of mental rotation. The conclusions about the effects of the test design could also be applied to other cognitive tests, where similar modifications and distinctions within trials are possible. It would be interesting, if similar more time-efficient tests are possible but also whether sex differences in performance can also be caused in other tests by varying the trials. Whereas a possible explanation related to spatial working memory, it is unclear if this only interacts with the mental rotation process, with tasks that in some way require spatial working memory or spatial abilities, or whether such sex differences by test design can be elicited in diverse cognitive tests just as, for example, the effects of gender stereotypes. Such an extension would require cognitive tests, which can be constructed in a similar multiple-choice design as the psychometric mental rotation tests.

While some possibilities to add further depth to the current studies and possible transfers and generalizations of practical relevance have been outlined, the same mechanisms and phenomena should also be investigated using different methods. The analyses of strategies, separate processing stages, cognitive effort, but also sex differences could be enhanced by further measures of neurological activity such as EEG or fNIRS.

10.7 Implications and Conclusions

The six studies presented here dealt with multiple topics involving mental rotation and provide some interesting insights regarding the link to physical activity, the test design, and sex differences. For a general link between simultaneous aerobic and cognitive activity, Dietrich and Audiffren (2011) propose an effect on cognitive performance depending on a distinction between the implicit and explicit system. We extended these results to a bidirectional interaction for explicit cognitive activity, which provides the possibility of extending the model. The theoretical prospect for both interference and facilitation in both directions could allow applications in both performance testing and training. Regarding the specific training of mental rotation, the link to manual rotation has proven to be effective to enhance performance (Adams et al., 2014; Wiedenbauer et al., 2007). In the fourth study, however, we identified the visualization as the main driver of improvements. This could allow more individual but possibly also cognitively and motorically less demanding training interventions. A point of theoretical concern regarding training effects is the mix of improvements by an intervention with improvements by repeated testing. In the third study, we demonstrated one approach using linear mixed models to better separate the combination of improvements.

Mental rotation training is not an end in itself, but is of particular interest because of sex differences, which are suspected to be one reason for sex differences in STEM attainment due to the link between spatial abilities and mathematical performance (Xie et al., 2020). One specific reason for sex differences in mental rotation performance could be the stimulus material. Traditionally, mental rotation tests use male stereotyped cube figures and stereotypes are known to influence performance (Steele & Aronson, 1995). In the fifth study, we identified a possible link between the stimulus material and performance on an individual level through the implicit attitudes. Exploring possibilities to modify the implicit attitudes could thus lead to reduced sex differences in mental rotation, in spatial abilities, and in STEM attainment.

As we have dealt with mental rotation performance, it is also important to consider how it is measured. As part of the second study, we have identified the exclusion of mirrored trials from analyses as a weakness and provided a solution through the incorporation of a second alternative. The advantage that mental rotation is useful in the solution of all trials was reflected in the monotonous relationship between angular disparity and reaction time in the results. Through the further comparison of trial designs in the sixth study, we identified further possible modifications of mental rotation tests. These provide opportunities to isolate what is considered as mental rotation ability from other abilities involved in solving distractor trials but also an increased statistical power. A complementary approach to enhance our understanding of the mental rotation ability is the further analysis of manual rotation by separating the phases as in the fourth study. As conclusions about mental rotation rely on test performance, it is important to understand the different abilities and phases involved and these should be further investigated. The results of the first study regarding cognitive effort and broad neurological resource conflicts could, on the other hand, help identify limits of cognitive testing aside from known speed-accuracy trade-offs. Another aspect identified in the sixth study regarding the trial design is that it could be a possible reason for varying sex differences in mental rotation performance. This again underlines the importance of understanding which abilities are involved and measured in a test.

In conclusion, this thesis provides multiple results that can help testing, understanding, and improving mental rotation ability. These can mainly improve and inspire future research on the topics, but the practical importance due to the link between mental rotation and mathematical abilities should not be neglected.

11 Declarations

11.1 Ethical Standards

All experiments were conducted according to ethical declaration of Helsinki. I communicated all considerations necessary to assess the question of ethical legitimacy of the studies.

11.2 Informed Consent to Participate and to Publish

Informed consent to participate and to publish anonymous results was obtained from all individual participants included in all studies.

11.3 Acknowledgements

I want to thank the following members of the University of Regensburg for their specific support in some of the studies. For their support with the data collection, I thank Arne Engelhardt (first, second, and fourth study) and Anna Wargel (fourth study). For his help with programming the experiment, I thank Alexander Kalus (fifth study). Moreover, I thank Philipp Hofmann and Markus Siebertz for their support with testing the stimulus library (second study) and Markus Siebertz also for helpful discussions regarding Bayesian statistics (fourth study).

11.4 Open Research Practices

The data that support the findings of the studies are available at https://osf.io/2m6wn/ (first study), https://osf.io/dr9mv/ (second study), https://osf.io/hxa2s/ (fourth study), https://osf.io/tdhvq/ (fifth study), https://osf.io/56fk3/ (sixth study).

The code to conduct the experiments and perform the analyses are available at https://osf.io/2m6wn/ (experiment of first study), https://github.com/LeonardoJost/MRExperiment (experiment of second study), https://github.com/LeonardoJost/MRlibrary (stimulus library of second study), https://github.com/LeonardoJost/TimeAnalysis (analysis of third study), https://github.com/LeonardoJost/MMR (experiment and analysis of fourth study), https://github.com/LeonardoJost/MCMR (experiment and analysis of sixth study).

The first, second, and third study were not preregistered. The fourth, fifth, and sixth study were preregistered at https://osf.io/xz3ma, https://osf.io/z9reu, and https://osf.io/b78yx, respectively.

11.5 Competing Interests

There were no competing interests for any of the studies.

11.6 Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied*, *13*(2), 57–78. https://doi.org/10.1037/1076-898X.13.2.57

Adams, D. M., Stull, A. T., & Hegarty, M. (2014). Effects of Mental and Manual Rotation Training on Mental and Manual Rotation Performance. *Spatial Cognition and Computation*, *14*(3), 169–198. https://doi.org/10.1080/13875868.2014.913050

Alexander, G. M., & Evardone, M. (2008). Blocks and bodies: Sex differences in a novel version of the Mental Rotations Test. *Hormones and Behavior*, *53*(1), 177–184. https://doi.org/10.1016/j.yhbeh.2007.09.014

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 1–7. https://doi.org/10.1111/bmsp.12004

Arts, F. J. P., & Kuipers, H. (1994). The relation between power output, oxygen uptake and heart rate in male athletes. *International Journal of Sports Medicine*, *15*(5), 228–231. https://doi.org/10.1055/s-2007-1021051

Au, J., Gibson, B. C., Bunarjo, K., Buschkuehl, M., & Jaeggi, S. M. (2020). Quantifying the Difference Between Active and Passive Control Groups in Cognitive Interventions Using Two Meta-analytical Approaches. *Journal of Cognitive Enhancement*, *4*(2), 192–210. https://doi.org/10.1007/s41465-020-00164-6

Audiffren, M., & André, N. (2015). The strength model of self-control revisited: Linking acute and chronic effects of exercise on executive functions. *Journal of Sport and Health Science*, *4*(1), 30–

46. https://doi.org/10.1016/j.jshs.2014.09.002

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Alday, P., Kleinschmidt, D., José Bayoán Santiago Calderón, P., Zhan, L., Noack, A., Arslan, A., Bouchet-Valat, M., Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P. K., Babayan, S., & Gagnon, Y. L. (2021). *JuliaStats/MixedModels.jl: v4.0.0*. Zenodo. https://doi.org/10.5281/zenodo.5111017

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967*. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Battista, C., & Peters, M. (2010). Ecological aspects of mental rotation around the vertical and horizontal axis. *Journal of Individual Differences*, *31*(2), 110–113. https://doi.org/10.1027/1614-0001/a000020

Bauer, R., Jost, L., Günther, B., & Jansen, P. (2021). Pupillometry as a measure of cognitive load in mental rotation tasks with abstract and embodied figures. *Psychological Research*, *0123456789*. https://doi.org/10.1007/s00426-021-01568-5

Bauer, R., Jost, L., & Jansen, P. (2021). The effect of mindfulness and stereotype threat in mental rotation: a pupillometry study. *Journal of Cognitive Psychology*, *0*(0), 1–16. https://doi.org/10.1080/20445911.2021.1967366

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. https://doi.org/10.1037/0022-3514.74.5.1252

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*(6), 351–355. https://doi.org/10.1111/j.1467-8721.2007.00534.x

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of psychophysiology, 2nd ed.* (pp. 142–162). Cambridge University Press.

Bentley, D. J., Newell, J., & Bishop, D. (2007). Incremental Exercise Test Design and Analysis. *Sports Medicine*, *37*(7), 575–586. https://doi.org/10.2165/00007256-200737070-00002

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Bishop, P. A., & Herron, R. L. (2015). Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International Journal of Exercise Science*, *8*(3), 297–302. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833473/

Boone, A. P., & Hegarty, M. (2017). Sex differences in mental rotation tasks: Not just in the mental rotation process! *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(7), 1005–1019. https://doi.org/10.1037/xlm0000370

Bors, D. A., & Vigneau, F. (2011). Sex differences on the mental rotation test: An analysis of item types. *Learning and Individual Differences*, *21*(1), 129–132. https://doi.org/10.1016/j.lindif.2010.09.014

Borst, G., Standing, G., & Kosslyn, S. M. (2012). Fear and anxiety modulate mental rotation. *Journal* *of Cognitive Psychology*, *24*(6), 665–671. https://doi.org/10.1080/20445911.2012.679924

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, *2*(1), 1–38. https://doi.org/10.5334/joc.72

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Buckley, J., Seery, N., & Canty, D. (2018). A Heuristic Framework of Spatial Ability: a Review and Synthesis of Spatial Factor Literature to Support its Translation into STEM Education. *Educational Psychology Review*, *30*(3), 947–972. https://doi.org/10.1007/s10648-018-9432-z

Büsch, D., Pabst, J., Naundorf, F., Braun, J., Marschall, F., Schumacher, K., Wilhelm, A., & Granacher, U. (2015). Subjektive Beanspruchung im Krafttraining. In U. Granacher (Ed.), *Krafttraining: "Kraftvoll durchs Leben": Jahrestagung der dvs-Sektion Trainingswissenschaft vom 28.-30. Mai 2015 in Potsdam (Abstractband)* (p. 13). Potsdam: Uni-Print.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Caldwell, A., & Lakens, D. (2019). *Power Analysis with Superpower*. https://aaroncaldwell.us/SuperpowerBook/

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations With Behavior and Explicit

Attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Campbell, M. J., Toth, A. J., & Brady, N. (2018). Illuminating sex differences in mental rotation using pupillometry. *Biological Psychology*, *138*(February), 19–26. https://doi.org/10.1016/j.biopsycho.2018.08.003

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W., & Thulborn, K. (1999). Graded functional activation in the visuospatial system with the amount of task demand. *Journal of Cognitive Neuroscience*, *11*(1), 9–24. https://doi.org/10.1162/089892999563210

Chang, Y. K., Labban, J. D., Gapin, J. I., & Etnier, J. L. (2012). The effects of acute exercise on cognitive performance: A meta-analysis. *Brain Research*, *1453*(250), 87–101. https://doi.org/10.1016/j.brainres.2012.02.068

Coast, J. R., & Welch, H. G. (1985). Linear increase in optimal pedal rate with increased power output in cycle ergometry. *European Journal of Applied Physiology and Occupational Physiology*, *53*(4), 339–342. https://doi.org/10.1007/BF00422850

Coyle, E. F., & González-Alonso, J. (2001). Cardiovascular Drift during Prolonged Exercise: New Perspectives. *Exercise and Sport Sciences Reviews*, *29*(2), 88–92. https://doi.org/10.1097/00003677-200104000-00009

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit Measures: A Normative Analysis and Review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

Debarnot, U., Piolino, P., Baron, J. C., & Guillot, A. (2013). Mental Rotation: Effects of Gender, Training and Sleep Consolidation. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0060296

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference.*

Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dietrich, A., & Audiffren, M. (2011). The reticular-activating hypofrontality (RAH) model of acute exercise. *Neuroscience & Biobehavioral Reviews*, *35*(6), 1305–1325. https://doi.org/10.1016/j.neubiorev.2011.02.001

Dodwell, G., Müller, H. J., & Töllner, T. (2019). Electroencephalographic evidence for improved visual working memory performance during standing and exercise. *British Journal of Psychology*, *110*(2), 400–427. https://doi.org/10.1111/bjop.12352

Doyle, R. A., & Voyer, D. (2013). Bodies and occlusion: Item types, cognitive processes, and gender differences in mental rotation. *Quarterly Journal of Experimental Psychology*, *66*(4), 801–815. https://doi.org/10.1080/17470218.2012.719529

Doyle, R. A., & Voyer, D. (2018). Photographs of real human figures: Item types and persistent sex differences in mental rotation. *Quarterly Journal of Experimental Psychology*, *71*(11), 2411–2420. https://doi.org/10.1177/1747021817742079

Eder, A. B., Leuthold, H., Rothermund, K., & Schweinberger, S. R. (2012). Automatic response activation in sequential affective priming: An ERP study. *Social Cognitive and Affective Neuroscience*, *7*(4), 436–445. https://doi.org/10.1093/scan/nsr033

Englert, C. (2016). The strength model of self-control in sport and exercise psychology. *Frontiers in Psychology*, *7*(MAR), 1–9. https://doi.org/10.3389/fpsyg.2016.00314

Faria, E. W., Parker, D. L., & Faria, I. E. (2005a). The Science of Cycling: Factors affecting Performance - Part 2. In *Sports Medicine* (Vol. 35, Issue 4, pp. 313–337). https://doi.org/10.2165/00007256-200535040-00002

Faria, E. W., Parker, D. L., & Faria, I. E. (2005b). The Science of Cycling: Physiology and Training

– Part 1. *Sports Medicine*, *35*(4), 313–337. https://doi.org/10.2165/00007256-200535040-00003

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

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*(2), 115–141. https://doi.org/10.1080/02699930125908

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobstrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037//0022-3514.69.6.1013

Feingold, A. (2009). Effect Sizes for Growth-Modeling Analysis for Controlled Clinical Trials in the Same Metric as for Classical Analysis. *Psychological Methods*, *14*(1), 43–53. https://doi.org/10.1037/a0014699.Effect

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fernández-Méndez, L. M., Contreras, M. J., & Elosúa, M. R. (2018). From What Age Is Mental Rotation Training Effective? Differences in Preschool Age but Not in Sex. *Frontiers in Psychology*, *9*(MAY), 1–10. https://doi.org/10.3389/fpsyg.2018.00753

Fisher, M. L., Meredith, T., & Gray, M. (2018). Sex Differences in Mental Rotation Ability Are a Consequence of Procedure and Artificiality of Stimuli. *Evolutionary Psychological Science*, *4*(2), 124–133. https://doi.org/10.1007/s40806-017-0120-x

Foulkes, D., & Hollifield, M. (1989). Responses to picture-plane and depth mental-rotation stimuli in children and adults. *Bulletin of the Psychonomic Society*, *27*(4), 327–330. https://doi.org/10.3758/BF03334617

Gardony, A. L., Taylor, H. A., & Brunyé, T. T. (2014). What Does Physical Rotation Reveal About Mental Rotation? *Psychological Science*, *25*(2), 605–612. https://doi.org/10.1177/0956797613503174

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

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *1*(1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

Green, P., & MacLeod, C. J. (2016). SIMR : an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Guilford, J. P., & Zimmerman, W. S. (1948). The Guilford-Zimmerman Aptitude Survey. *Journal of Applied Psychology*, *32*(1), 24.

Guizzo, F., Moè, A., Cadinu, M., & Bertolli, C. (2019). The role of implicit gender spatial stereotyping in mental rotation performance. *Acta Psychologica*, *194*(September 2018), 63–68. https://doi.org/10.1016/j.actpsy.2019.01.013

Gyurak, A., Gross, J. J., & Etkin, A. (2011). Explicit and implicit emotion regulation: A dual-process framework. *Cognition & Emotion*, *25*(3), 400–412. https://doi.org/10.1080/02699931.2010.544160

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press, Taylor & Francis.

Halpern, D. F., & Tan, U. (2001). Stereotypes and steroids: Using a psychobiosocial model to understand cognitive sex differences. *Brain and Cognition*, *45*(3), 392–414. https://doi.org/10.1006/brcg.2001.1287

Hausmann, M. (2014). Arts versus science - Academic background implicitly activates gender stereotypes on cognitive abilities with threat raising men's (but lowering women's) performance. *Intelligence*, *46*(1), 235–245. https://doi.org/10.1016/j.intell.2014.07.004

Hausmann, M., Schoofs, D., Rosenthal, H. E. S., & Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences-A psychobiosocial approach. *Psychoneuroendocrinology*, *34*(3), 389–401. https://doi.org/10.1016/j.psyneuen.2008.09.019

Hausmann, M., Slabbekoorn, D., Van Goozen, S. H. M., Cohen-Kettenis, P. T., & Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. *Behavioral Neuroscience*, *114*(6), 1245–1250. https://doi.org/10.1037/0735-7044.114.6.1245

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hegarty, M. (2018). Ability and sex differences in spatial thinking: What does the mental rotation test really measure? *Psychonomic Bulletin and Review*, *25*(3), 1212–1219. https://doi.org/10.3758/s13423-017-1347-z

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 *The Cambridge Handbook of Visuospatial Thinking* (pp. 121–169). Cambridge University Press. https://doi.org/10.1017/CBO9780511610448.005

Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones?

Quarterly Journal of Experimental Psychology, *61*(5), 683–689. https://doi.org/10.1080/17470210701822967

Heil, M., Jansen, P., Quaiser-Pohl, C., & Neuburger, S. (2012). Gender-specific effects of artificially induced gender beliefs in mental rotation. *Learning and Individual Differences*, *22*(3), 350–353. https://doi.org/10.1016/j.lindif.2012.01.004

Heil, M., & Rolke, B. (2002). Toward a chronopsychophysiology of mental rotation. *Psychophysiology*, *39*(4), 414–422. https://doi.org/10.1111/1469-8986.3940414

Heil, M., Rösler, F., Link, M., & Bajric, J. (1998). What is improved if a mental rotation task is repeated - The efficiency of memory access, or the speed of a transformation routine? *Psychological Research*, *61*(2), 99–106. https://doi.org/10.1007/s004260050016

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *26*(1), 101–119. https://doi.org/10.4473/TPM26.1.6

Hohmann, A., Lames, M., & Letzelter, M. (2002). *Einführung in die Trainingswissenschaft*. Wiebelsheim: Limpert. https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/index/index/docId/34678

Hopkins, W. G., Schabort, E. J., & Hawley, J. A. (2001). Reliability of power in physical performance tests. In *Sports Medicine* (Vol. 31, Issue 3, pp. 211–234). Adis International Ltd. https://doi.org/10.2165/00007256-200131030-00005

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

Hutcherson, C. A., Seppala, E. M., & Gross, J. J. (2008). Loving-Kindness Meditation Increases Social Connectedness. *Emotion*, *8*(5), 720–724. https://doi.org/10.1037/a0013237

Hyun, J. S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. *Psychonomic Bulletin and Review*, *14*(1), 154–158. https://doi.org/10.3758/BF03194043

Ilan, A. B., & Miller, J. (1994). A Violation of Pure Insertion: Mental Rotation and Choice Reaction Time. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(3), 520–536. https://doi.org/10.1037/0096-1523.20.3.520

Janczyk, M., Pfister, R., Crognale, M. A., & Kunde, W. (2012). Effective rotations: Action effects determine the interplay of mental and manual rotations. *Journal of Experimental Psychology: General*, *141*(3), 489–501. https://doi.org/10.1037/a0026997

Jansen-Osmann, P., & Heil, M. (2007a). Maintaining readiness for mental rotation interferes with perceptual processes in children but with response selection in adults. *Acta Psychologica*, *126*(3), 155–168. https://doi.org/10.1016/j.actpsy.2006.11.005

Jansen-Osmann, P., & Heil, M. (2007b). 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

Jansen, P., & Heil, M. (2009). Gender Differences in Mental Rotation Across Adulthood. *Experimental Aging Research*, *36*(1), 94–104. https://doi.org/10.1080/03610730903422762

Jansen, P., & Lehmann, J. (2013). Mental rotation performance in soccer players and gymnasts in an object-based mental rotation task. *Advances in Cognitive Psychology*, *9*(2), 92–98. https://doi.org/10.2478/vl0053-008-0135-8

Jolicœur, P., Regehr, S., Smith, L. B. J. P., & Smith, G. N. (1985). Mental rotation of representations of two-dimensional and three-dimensional objects. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *39*(1), 100–129. https://doi.org/10.1037/h0080118

Jordan, K., Wüstenberg, T., Heinze, H. J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia*, *40*(13), 2397–2408. https://doi.org/10.1016/S0028-3932(02)00076-3

Jost, L., & Jansen, P. (2020). A novel approach to analyzing all trials in chronometric mental rotation and description of a flexible extended library of stimuli. *Spatial Cognition & Computation*, *20*(3), 234–256. https://doi.org/10.1080/13875868.2020.1754833

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, *79*(1, Pt.1), 164–167. https://doi.org/10.1037/h0026952

Kaltner, S., & Jansen, P. (2014). Emotion and affect in mental imagery: do fear and anxiety manipulate mental rotation performance? *Frontiers in Psychology*, *5*, 792. https://doi.org/10.3389/fpsyg.2014.00792

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

Kennedy, D. O., & Scholey, A. B. (2000). Glucose administration, heart rate and cognitive performance: Effects of increasing mental effort. *Psychopharmacology*, *149*(1), 63–71. https://doi.org/10.1007/s002139900335

Kerkman, D. D., Wise, J. C., & Harwood, E. A. (2000). Impossible "mental rotation" problems: A mismeasure of women's spatial abilities? *Learning and Individual Differences*, *12*(3), 253–269. https://doi.org/10.1016/S1041-6080(01)00039-5

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Khooshabeh, P., & Hegarty, M. (2010). Representations of shape during mental rotation. *AAAI Spring Symposium - Technical Report*, *SS-10-02*, 15–20.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory and Cognition*, *29*(5), 745–756. https://doi.org/10.3758/BF03200477

Krüger, J. K., & Suchan, B. (2016). You should be the specialist! weak mental rotation performance in aviation security screeners - reduced performance level in aviation security with no gender effect. *Frontiers in Psychology*, *7*(MAR), 333. https://doi.org/10.3389/fpsyg.2016.00333

Krüger, M., & Krist, H. (2009). Imagery and motor processes - When are they connected? The mental rotation of body parts in development. *Journal of Cognition and Development*, *10*(4), 239–261. https://doi.org/10.1080/15248370903389341

Lambourne, K., & Tomporowski, P. (2010). The effect of exercise-induced arousal on cognitive task performance: A meta-regression analysis. *Brain Research*, *1341*, 12–24. https://doi.org/10.1016/j.brainres.2010.03.091

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), 1–14. https://doi.org/10.1371/journal.pone.0130834

Larson, C. L., Aronoff, J., & Steuer, E. L. (2012). Simple geometric shapes are implicitly associated with affective value. *Motivation and Emotion*, *36*(3), 404–413. https://doi.org/10.1007/s11031-011-9249-2

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. *ArXiv:1405.2094*. http://arxiv.org/abs/1405.2094

Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, *2*(2), 155–162. https://doi.org/10.1111/j.2041-210X.2010.00063.x

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, *51*(1), 40–60.

https://doi.org/10.3758/s13428-018-1076-x

Liesefeld, H. R., & Zimmer, H. D. (2013). Think spatial: The representation in mental rotation is nonvisual. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(1), 167–182. https://doi.org/10.1037/a0028904

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, *44*, 187–192.

Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability : A Meta-Analysis. *Child Development*, *56*(6), 1479–1498. https://doi.org/10.2307/1130467

MacIntosh, B. R., Neptune, R. R., & Horton, J. F. (2000). Cadence, power, and muscle activation in cycle ergometry. *Medicine and Science in Sports and Exercise*, *32*(7), 1281–1287. https://doi.org/10.1097/00005768-200007000-00015

Mammarella, N. (2011). Is there a "special relationship" between unconscious emotions and visual imagery? Evidence from a mental rotation test. *Consciousness and Cognition*, *20*(2), 444–448. https://doi.org/10.1016/j.concog.2010.10.012

Martin, K., Meeusen, R., Thompson, K. G., Keegan, R., & Rattray, B. (2018). Mental Fatigue Impairs Endurance Performance: A Physiological Explanation. *Sports Medicine*, *48*(9), 2041–2051. https://doi.org/10.1007/s40279-018-0946-9

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. In *Behavior Research Methods* (Vol. 44, Issue 2, pp. 314–324). https://doi.org/10.3758/s13428-011-0168-7

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Meneghetti, C., Carbone, E., Di Maggio, A., Toffalini, E., & Borella, E. (2018). Mental rotation training in older adults: The role of practice and strategy. *Psychology and Aging*, *33*(5), 814–831. https://doi.org/10.1037/pag0000275

Meneghetti, C., Cardillo, R., Mammarella, I. C., Caviola, S., & Borella, E. (2017). The role of practice and strategy in mental rotation training: transfer and maintenance effects. *Psychological Research*, *81*(2), 415–431. https://doi.org/10.1007/s00426-016-0749-2

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Mix, K. S., Hambrick, D. Z., Satyam, V. R., Burgoyne, A. P., & Levine, S. C. (2018). The latent structure of spatial skill: A test of the 2 × 2 typology. *Cognition*, *180*, 268–278. https://doi.org/https://doi.org/10.1016/j.cognition.2018.07.012

Moè, A. (2016). Does experience with spatial school subjects favour girls' mental rotation performance? *Learning and Individual Differences*, *47*, 11–16. https://doi.org/10.1016/j.lindif.2015.12.007

Moè, A. (2018). Effects of Group Gender Composition on Mental Rotation Test Performance in Women. *Archives of Sexual Behavior*, *47*(8), 2299–2305. https://doi.org/10.1007/s10508-018-1245-0

Moè, A., Jansen, P., & Pietsch, S. (2018). Childhood preference for spatial toys. Gender differences and relationships with mental rotation in STEM and non-STEM students. *Learning and Individual Differences*, *68*, 108–115. https://doi.org/10.1016/j.lindif.2018.10.003

Moè, A., & Pazzaglia, F. (2006). Following the instructions!. Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, *16*(4), 369–377. https://doi.org/10.1016/j.lindif.2007.01.002

Monahan, J. S., Harke, M. A., & Shelley, J. R. (2008). Computerizing the mental rotations test: Are gender differences maintained? *Behavior Research Methods*, *40*(2), 422–427. https://doi.org/10.3758/BRM.40.2.422

Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, *43*(9), 1–14.

Neubauer, A. C., & Hofer, G. (2020). Self-estimates of abilities are a better reflection of individuals' personality traits than of their abilities and are also strong predictors of professional interests. *Personality and Individual Differences*, 109850. https://doi.org/10.1016/j.paid.2020.109850

Newcombe, N. S., & Shipley, T. F. (2015). Thinking About Spatial Thinking: New Typology, New Assessments. In J. S. Gero (Ed.), *Studying Visual and Spatial Reasoning for Design Creativity* (pp. 179–192). Springer Netherlands. https://doi.org/10.1007/978-94-017-9297-4_10

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Palumbo, L., Ruta, N., & Bertamini, M. (2015). Comparing Angular and Curved Shapes in Terms of Implicit Associations and Approach/Avoidance Responses. *PLOS ONE*, *10*(10), e0140043. https://doi.org/10.1371/journal.pone.0140043

Paschke, K., Jordan, K., Wüstenberg, T., Baudewig, J., & Leo Müller, J. (2012). Mirrored or identical - Is the role of visual perception underestimated in the mental rotation process of

3D-objects?: A combined fMRI-eye tracking-study. *Neuropsychologia*, *50*(8), 1844–1851. https://doi.org/10.1016/j.neuropsychologia.2012.04.010

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition*, *57*(2), 176–184. https://doi.org/10.1016/j.bandc.2004.08.052

Peters, M., & Battista, C. (2008). Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. *Brain and Cognition*, *66*(3), 260–264. https://doi.org/10.1016/j.bandc.2007.09.003

Peters, M., Chisholm, P., & Laeng, B. (1995). Spatial Ability, Student Gender, and Academic Performance. *Journal of Engineering Education*, *84*(1), 69–73. https://doi.org/10.1002/j.2168-9830.1995.tb00148.x

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., Manning, J. T., & Reimers, S. (2007). The effects of sex, sexual orientation, and digit ratio (2D:4D) on mental rotation performance. *Archives of Sexual Behavior*, *36*(2), 251–260. https://doi.org/10.1007/s10508-006-9166-8

Piaget, & Inhelder. (1956). *The child's conception of space*. London: Routledge and Kegan Paul.

Quaiser-Pohl, C., & Lehmann, W. (2002). Girls' spatial abilities: Charting the contributions of experiences and attitudes in different academic groups. *British Journal of Educational Psychology*, *72*(2), 245–260. https://doi.org/10.1348/000709902158874

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rahe, M., & Quaiser-Pohl, C. (2020). Cubes or pellets in mental-rotation tests: Effects on gender differences and on the performance in a subsequent math test. *Behavioral Sciences*, *10*(1), 8–15. https://doi.org/10.3390/bs10010012

Rahe, M., Ruthsatz, V., Jansen, P., & Quaiser-Pohl, C. (2019). Different practice effects for males and females by psychometric and chronometric mental-rotation tests. *Journal of Cognitive Psychology*, *31*(1), 92–103. https://doi.org/10.1080/20445911.2018.1561702

Rahe, M., Ruthsatz, V., & Quaiser-Pohl, C. (2020). Influence of the stimulus material on gender differences in a mental-rotation test. *Psychological Research*, *0123456789*. https://doi.org/10.1007/s00426-020-01450-w

Rahe, M., Ruthsatz, V., Schürmann, L., & Quaiser-Pohl, C. (2019). The effects of feedback on the gender differences in the performance in a chronometric mental-rotation test. *Journal of Cognitive Psychology*, *31*(4), 467–475. https://doi.org/10.1080/20445911.2019.1621872

Riby, L., Perfect, T., & Stollery, B. (2004). The effects of age and task domain on dual task performance: A meta-analysis. *European Journal of Cognitive Psychology*, *16*(6), 863–891. https://doi.org/10.1080/09541440340000402

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

Robert, M., & Chevrier, E. (2003). Does men's advantage in mental rotation persist when real three-dimensional objects are either felt or seen? *Memory and Cognition*, *31*(7), 1136–1145.

https://doi.org/10.3758/BF03196134

Ruthsatz, V., Neuburger, S., Jansen, P., & Quaiser-Pohl, C. (2014). Pellet Figures, the Feminine Answer to Cube Figures? Influence of Stimulus Features and Rotational Axis on the Mental-Rotation Performance of Fourth-Grade Boys and Girls. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8684 LNAI* (pp. 370–382). https://doi.org/10.1007/978-3-319-11215-2_26

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., Neuburger, S., Rahe, M., Jansen, P., & Quaiser-Pohl, C. (2017). The gender effect in 3D-Mental-rotation performance with familiar and gender-stereotyped objects–a study with elementary school children. *Journal of Cognitive Psychology*, *29*(6), 717–730. https://doi.org/10.1080/20445911.2017.1312689

Schaefer, S., Jagenow, D., Verrel, J., & Lindenberger, U. (2015). The influence of cognitive load and walking speed on gait regularity in children and young adults. *Gait and Posture*, *41*(1), 258–262. https://doi.org/10.1016/j.gaitpost.2014.10.013

Schaefer, S., & Schumacher, V. (2011). The interplay between cognitive and motor functioning in healthy Older adults: Findings from dual-task studies and suggestions for intervention. *Gerontology*, *57*(3), 239–246. https://doi.org/10.1159/000322197

Schäfer, S., Huxhold, O., & Lindenberger, U. (2006). Healthy mind in healthy body? A review of sensorimotor-cognitive interdependencies in old age. *European Review of Aging and Physical Activity*, *3*(2), 45–54. https://doi.org/10.1007/s11556-006-0007-5

Scheer, C., Maturana, F. M., & Jansen, P. (2018). Sex differences in a chronometric mental rotation test with cube figures: A behavioral, electroencephalography, and eye-tracking pilot study.

NeuroReport, *29*(10), 870–875. https://doi.org/10.1097/WNR.0000000000001046

Selarka, D., Rosenbaum, R. S., Lapp, L., & Levine, B. (2019). Association between self-reported and performance-based navigational ability using internet-based remote spatial memory assessment. *Memory*, *27*(5), 723–728. https://doi.org/10.1080/09658211.2018.1554082

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, *93*(3), 604–614. https://doi.org/10.1037/0022-0663.93.3.604

Shepard, R. N., & Cooper, L. (1986). Mental images and their transformations. In *Mental images and their transformations*. The MIT Press.

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

Shepard, S., & Metzler, D. (1988). Mental Rotation: Effects of Dimensionality of Objects and Type of Task. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 3–11. https://doi.org/10.1037/0096-1523.14.1.3

Shiffrar, M. M., & Shepard, R. N. (1991). Comparison of cube rotations around axes inclined relative to the environment or to the cube. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 44–54. https://doi.org/10.1037//0096-1523.17.1.44

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Spruyt, A., Hermans, D., De Houwer, J., & Eelen, P. (2002). On the nature of the affective priming effect: Affective priming of naming responses. *Social Cognition*, *20*(3), 227–256. https://doi.org/10.1521/soco.20.3.227.21106

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior*

Research Methods, Instruments, & Computers, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Such, U., & Meyer, T. (2010). Die maximale herzfrequenz. *Deutsche Zeitschrift Fur Sportmedizin*, *61*(12), 310–311.

Titze, C., Heil, M., & Jansen, P. (2008). Gender Differences in the Mental Rotations Test (MRT) Are Not Due to Task Complexity. *Journal of Individual Differences*, *29*(3), 130–133. https://doi.org/10.1027/1614-0001.29.3.130

Titze, C., Heil, M., & Jansen, P. (2010). Pairwise presentation of cube figures does not reduce gender differences in mental rotation performance. *Journal of Individual Differences*, *31*(2), 101–105. https://doi.org/10.1027/1614-0001/a000018

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. https://doi.org/10.1037/a0028446

Van Cutsem, J., Marcora, S., De Pauw, K., Bailey, S., Meeusen, R., & Roelands, B. (2017). The Effects of Mental Fatigue on Physical Performance: A Systematic Review. *Sports Medicine*, *47*(8), 1569–1588. https://doi.org/10.1007/s40279-016-0672-0

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

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

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin and Review*, *18*(2), 267–277. https://doi.org/10.3758/s13423-010-0042-0

Voyer, D., Butler, T., Cordero, J., Brake, B., Silbersweig, D., Stern, E., & Imperato-McGinley, J. (2006). The relation between computerized and paper-and-pencil mental rotation tasks: A validation study. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 928–939. https://doi.org/10.1080/13803390591004310

Voyer, D., & Hou, J. (2006). Type of items and the magnitude of gender differences on the mental rotations test. In *Canadian Journal of Experimental Psychology* (Vol. 60, Issue 2, pp. 91–100). https://doi.org/10.1037/cjep2006010

Voyer, D., & Jansen, P. (2016). Sex differences in chronometric mental rotation with human bodies. *Psychological Research*, *80*(6), 974–984. https://doi.org/10.1007/s00426-015-0701-x

Voyer, D., Jansen, P., & Kaltner, S. (2017). Mental rotation with egocentric and object-based transformations. *Quarterly Journal of Experimental Psychology*, *70*(11), 2319–2330. https://doi.org/10.1080/17470218.2016.1233571

Voyer, D., Nolan, C., & Voyer, S. (2000). The relation between experience and spatial performance in men and women. *Sex Roles*, *43*(11–12), 891–915. https://doi.org/https://doi.org/10.1023/A:1011041006679

Voyer, D., Saint-Aubin, J., Altman, K., & Doyle, R. A. (2020). Sex differences in tests of mental rotation: Direct manipulation of strategies with eye-tracking. *Journal of Experimental Psychology:*

Human Perception and Performance, *46*(9), 871–889. https://doi.org/10.1037/xhp0000752

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

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin and Review*, *24*(2), 307–334. https://doi.org/10.3758/s13423-016-1085-7

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. In *Psychonomic Bulletin and Review* (Vol. 14, Issue 5, pp. 779–804). Psychonomic Society Inc. https://doi.org/10.3758/BF03194105

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

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2010). Accomplishment in Science, Technology, Engineering, and Mathematics (STEM) and Its Relation to STEM Educational Dose: A 25-Year Longitudinal Study. *Journal of Educational Psychology*, *102*(4), 860–871. https://doi.org/10.1037/a0019454

Walton, G. M., & Cohen, G. L. (2003). Stereotype Lift. *Journal of Experimental Social Psychology*, *39*(5), 456–467. https://doi.org/10.1016/S0022-1031(03)00019-2

Wang, Y., & Zhang, Q. (2016). Affective priming by simple geometric shapes: Evidence from event-related brain Potentials. *Frontiers in Psychology*, *7*(JUN). https://doi.org/10.3389/fpsyg.2016.00917

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, *68*(1), 77–94. https://doi.org/10.1016/S0010-0277(98)00032-8

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wiedenbauer, G., & Jansen-Osmann, P. (2007). Mental Rotation Ability of Children with Spina Bifida: What Influence Does Manual Rotation Training Have? *Developmental Neuropsychology*, *32*(3), 809–824. https://doi.org/10.1080/87565640701539626

Wiedenbauer, G., & Jansen-Osmann, P. (2008). Manual training of mental rotation in children. *Learning and Instruction*, *18*(1), 30–41. https://doi.org/10.1016/j.learninstruc.2006.09.009

Wiedenbauer, G., Schmid, J., & Jansen-Osmann, P. (2007). Manual training of mental rotation. *European Journal of Cognitive Psychology*, *19*(1), 17–36. https://doi.org/10.1080/09541440600709906

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *ArXiv:1308.5499*. http://arxiv.org/abs/1308.5499

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, *1*(1), 7–18. https://doi.org/10.1093/jole/lzv003

Witkin, H. A. (1971). *A manual for the embedded figures tests*. Consulting Psychologists Press.

Wohlschläger, A. (2001). Mental object rotation and the planning of hand movements. *Perception and Psychophysics*, *63*(4), 709–718. https://doi.org/10.3758/BF03194431

Wohlschläger, A., & Wohlschläger, A. (1998). Mental and Manual Rotation. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 397–412. https://doi.org/10.1037/0096-1523.24.2.397

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin and Review*, *15*(4), 763–771. https://doi.org/10.3758/PBR.15.4.763

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

Xue, J., Li, C., Quan, C., Lu, Y., Yue, J., & Zhang, C. (2017). Uncovering the cognitive processes underlying mental rotation: An eye-movement study. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-10683-6

Zacks, J., Mires, J., Tversky, B., & Hazeltine, E. (2000). Mental spatial transformations of objects and perspective. *Spatial Cognition and Computation*, *2*(4), 315–332. https://doi.org/10.1023/A:1015584100204

Appendix

Comparing BIC-Based Bayes Factors to a Model Generation Using Likelihood Ratio Tests

For the random effects structure of the linear mixed model we have mainly followed the approach of Bates et al. (2015), who also showed that Bayesian modeling supports the same random effects structure. For the fixed effects, we have reduced the model using likelihood ratio tests (LRT) to eliminate non-significant effects at the significance level $\alpha = .05$. Assuming that it is sensible to use the same models for the computation of Bayes factors we get the following relationship between the p-values and Bayes factors:

For the comparison of two models m_1 and m_0, where m_0 is nested within m_1 (m_0 typically represents the null hypothesis and m_1 a model containing m_0 and additionally a fixed effect of interest), a LRT is based on

$$\chi^2(m_1, m_0) = -2(\log L_0 - \log L_1),$$

where L_i is the maximum likelihood for model m_i (and $\log L_\mathrm{i}$ is the logarithmic likelihood of the model). Wagenmakers (2007) derives an approximated Bayes factor (in favor of model m_0) for the comparison of these two models as

$$BF_{01} \approx \exp\left(\frac{\Delta \mathrm{BIC}_{10}}{2}\right),$$

Where BIC is the Bayesian information criterion and

$$\Delta \mathrm{BIC}_{10} = \mathrm{BIC}(m_1) - \mathrm{BIC}(m_0)$$

As

$$\mathrm{BIC}(m_1) = -2\log L_1 + \log(n) * \mathrm{df}(m_1)$$

BF_{01} can be computed by

$$2\log(BF_{01}) = \Delta \mathrm{BIC}_{10}$$

$$= \text{BIC}(m_1) - \text{BIC}(m_0)$$

$$= -2\log L_1 + \log(n) * \text{df}(m_1) - \left(-2\log L_0 + \log(n) * \text{df}(m_0)\right)$$

$$= -\chi^2(m_1, m_0) + \log(n)(\text{df}(m_1) - \text{df}(m_0))$$

or

$$BF_{01} = \exp((-\chi^2(m_1, m_0) + \log(n)\left(\Delta\text{df}(m_1, m_0)\right))/2).$$

Note that $\Delta\text{df}(m_1, m_0) = \text{df}(m_1) - \text{df}(m_0)$, which is the number of degrees of freedom introduced by the effect in question.

As n is fixed within an experiment, there is a monotonous relationship between BF_{01} and χ^2. As χ^2 is again monotonously related to p by the χ^2-distribution, there is a monotonous relationship between BF_{01} and p[18]. For the cutoffs of 3 (favoring the null hypothesis) or $\frac{1}{3}$ (favoring the alternative hypothesis) for the Bayes factor we can thus calculate for a given $\Delta\text{df}(m_1, m_0)$ for which values of χ^2 it is achieved. In the fourth study, $n = 121$ ("In such a hierarchical or multilevel design, it is not quite clear what n should be. In this case, the standard choice is to take n to be the number of subjects." Wagenmakers, 2007, p.798) and $\log(121) \approx 4.796$. Thus, for effects adding one degree of freedom (i.e., all dichotomous and numerical independent variables) we get that

$$BF_{01} < \frac{1}{3} \leftrightarrow \chi^2(m_1, m_0) > 6.993 \leftrightarrow p < .008$$

$$BF_{01} > 3 \leftrightarrow \chi^2(m_1, m_0) < 2.599 \leftrightarrow p > .107$$

For effects with two degrees of freedom (in the fourth study only the independent variable group and all interactions containing it) we get

[18] Similarly, a monotonous relationship between the Bayes factors using the calculator associated with Dienes (2008), the *t*-values, and the *p*-values emerges if comparable priors are used.

$$BF_{01} < \frac{1}{3} \leftrightarrow \chi^2(m_1, m_0) > 11.789 \leftrightarrow p < .003$$

$$BF_{01} > 3 \leftrightarrow \chi^2(m_1, m_0) < 7.395 \leftrightarrow p > .025$$

As a result, the Bayes factor strongly supports the null hypothesis in almost all cases where $\mathrm{p} > .05$ and the null hypothesis is not rejected under the frequentist approach. In the other direction, the Bayes factor deems some significant results as inconclusive and for two degrees of freedom even favors the null hypothesis. Indeed, this is somewhat in line with Wagenmakers (2007) who in figure 6 shows the possibility of a significant result under frequentist statistics and the Bayes factor favoring the null (see also the Jeffreys-Lindley paradox; Lindley, 1957).

Reviewing the model generation using the LRT where we eliminated non-significant effects, we can now compare it to a model generation using the Bayes factor with a threshold of 1 (i.e., always sticking with the more probable hypothesis). A unique threshold is needed, as effects can only be excluded or included and there can be no region where more data is needed (as for $\frac{1}{3} < BF_{01} < 3$). Using the approximation of Wagenmakers (2007) this is equivalent to choosing the lowest BIC as $\exp(0) = 1$. That is, if for any two models $BF_{01} < 1$ then $\Delta\mathrm{BIC}_{10} < 0$ and if $BF_{01} > 1$ then $\Delta\mathrm{BIC}_{10} > 0$. Choosing the model with lowest BIC is thus equivalent to removing all effects for which $BF_{01} > 1.$ Thus, the model selections by LRT or BIC (or AIC) are equivalent up to a different p-value cutoff α. While most are arbitrary, the choice of $BF = 1$ does distinguish between the most likely effects. For LRT, α can be chosen arbitrarily. The AIC is equivalent to an LRT-based model selection using $\alpha \approx 0.157$ (Matuschek et al., 2017). For the BIC the equivalent α depends on n. For a more thorough comparison see also Lewis et al. (2011).

Going back to the relationship between BF and p, we can get the following situation:

Assume an experiment and effects of one degree of freedom (i.e., either dichotomous or numerical). Assume an arbitrary significance level α (typically .05) and an arbitrary Bayes factor cutoff BF_0 in favor of the null hypothesis (typically 3). Due to the relationship between p and BF we can calculate an n_0, such that for all effects with $p > \alpha$ we get $BF > BF_0$ and for all effects with $p < \alpha$ we get $BF < BF_0$. Moreover, for all $n > n_0$ we get for all effects with $p > \alpha$ that $BF > BF_0$, while the other direction is maintained for all $n < n_0$.

For $\alpha = .05$ and $BF_0 = 3$ and as $\chi^2(1) \approx 3.841 \leftrightarrow p = .05$ we get

$$3 = \exp(-3.841 + \log(n_0))/2)$$

$$\leftrightarrow \log(n_0) = 2\log(3) + 3.841$$

$$\leftrightarrow n_0 \approx 419$$

(In the other direction for $BF_0 = \frac{1}{3}$ we get $n_0 = \exp\left(2\log\left(\frac{1}{3}\right) + 3.841\right) \approx 5$)

By loosely citing Dienes (2014) we can use the Bayes factor for the supporting interpretation of only the nonsignificant results. In these cases, $BF > BF_0$ and all nonsignificant effects support the null. On the other hand, all significant effects follow the frequentist interpretation that the alternative is true. As we are dealing with uncertain data or an uncertain world depending on statistical interpretation, this should not happen. We thus deem it necessary to interpret Bayes factors for all effects if they are used to aid interpretation of frequentist analyses.

As demonstrated by Wagenmakers (2007) or the Jeffreys-Lindley paradox, the conclusions of frequentist and Bayesian statistics might be divergent with increasing data. While frequentist favors rejecting the null, Bayesian will favor the null. By combining both theories a region of uncertainty is introduced, where both approaches disagree (assuming for simplicity some equivalency with frequentists not rejecting the null and Bayesians actually favoring the null). The size of this region can be easily calculated by the calculations above. Note that a disagreement typically occurs when Bayesians favor the null while frequentists favor the alternative. This applies to well-known examples against frequentist statistics: When n is large enough, anything will become significant. The opposite direction might theoretically occur for very small sample sizes (less than 5 in the example above).