Navigating Minds:

Exploring the Spatial Thinking Abilities of College Students

Monika Sweeney

Senior Thesis

April 30, 2024

Submitted in partial fulfillment of the requirements for the

Bachelor of Arts degree in Earth Science

_____

Advisor, Kirsten Menking

Abstract

Spatial thinking is a fundamental cognitive skill that is crucial for understanding concepts in Earth Science, yet challenging for many individuals to develop. This research aimed to investigate the impact of a Structural Geology course on students' spatial thinking abilities, examining potential disparities based on gender and expertise. Employing a longitudinal observational design, four spatial thinking assessments were administered to 11 students enrolled in Earth Science 271: Structural Geology at Vassar College. Repeated measures ANOVAs, conducted via a statistical analysis program (JASP), showed limited changes in test scores over the four-week period of observation. While gender and prior experience in Earth Science had little effect on performance, the study encountered challenges, such as a small sample size and short data collection period, which may have made it difficult to identify significant patterns. Future research could address these limitations by utilizing a larger and more diverse sample size and conducting observations over a longer period of time. This approach could improve findings' credibility and provide valuable insights for spatial thinking education strategies.

Table of Contents

Acknowledgements

This thesis would not have been possible without the guidance, support and wisdom of my thesis advisor, Professor Kirsten Menking, my major advisor, Professor Laura Haynes, and the Earth Science and Geography Department. Thank you also to my friends and family who have supported me throughout my time at Vassar!

I. Introduction

Spatial thinking is the cognitive process that allows one to visualize and manipulate objects by deriving meaning from characteristics such as size, orientation, or shape (K. Kastens & Passow, 2012). It encompasses the skills and knowledge that humans use to understand concepts of space and relies on the ability to mentally manipulate and understand spatial relationships, patterns, and configurations (Hegarty, 2010). Examples of spatial thinking skills include being able to visualize what a 3-D object looks like from a 2-D representation and being able to imagine what a shape would look like from another angle or if it was sliced apart (Uttal et al., 2013).

In Earth Science, spatial thinking is a valuable skill to possess (Kastens & Ishikawa, 2006). It is crucial for analyzing and interpreting geologic maps and to understand how subsurface rock units interact with landforms and surface features. Without an understanding of spatial relationships, Earth scientists would not be able to interpret cross-sections and stratigraphic columns of rock layers or reconstruct geologic history by correlating rock formations from different locations (McNeal & Petcovic, 2020).

Despite its importance, spatial thinking is a particularly challenging skill to master for both adolescents and adults (Ishikawa & Newcombe, 2021; Rebelsky, 1964). Research shows that one's spatial thinking abilities are influenced by a variety of factors such as genetics, brain development, environment, and experiences (King et al., 2019; Chow et al., 2013; Gauvain, 1992). Despite the difficulty people have in acquiring these skills, research has also found that with time and practice, spatial thinking skills can be developed and improved (Uttal et al., 2013; Sorby, 2007).

II. Literature Review

One of the founders of Geoscience Education Research, Kim Kastens, recognizes that spatial thinking is central to various subfields in Earth Science, such as mineralogy and plate tectonics. She believes that insights from spatial thinking could inform better teaching practices within Earth Science, while Earth Science could push the boundaries of spatial thinking (Kastens, 2021). According to Kastens and Passow (2015), "Spatial thinking is what we are doing when we derive meaning from the shape, size, orientation, position, direction or trajectory of objects, processes or phenomena, or the relative positions in the space of multiple objects, processes or phenomena." Spatial thinking involves spatial representations, such as maps; spatial concepts, such as direction; and spatial skills, such as envisioning what a mountain range looks like from different angles (Bednarz & Lee, 2011).

Significant moments in the field of spatial thinking date back as far as 600 BC with the astronomical and mathematical contributions of the Babylonians (Steele & Ganguli, 2022). They utilized their understanding of spatial relationships, specifically the positions and trajectory of celestial bodies relative to one another and to the Earth, to track planetary movement throughout the sky. From there, the Babylonians laid the foundations for the future of mathematics, navigation, and astronomy (Rochberg, 2002). More modern examples of spatial thinking include Alfred Wegener's Theory of Continental Drift, proposed in 1912. Wegener was able to visualize the continents once fitting together like jigsaw pieces and to imagine their movement across the globe over millions of years (Kastens & Passow, 2015).

Beyond Wegener's Theory of Continental Drift, spatial thinking is abundant in Earth science. Between 2008 and 2011, Kastens and her colleagues at Lamont-Doherty Earth Observatory analyzed 11 NYS Earth Science Regents exams. They found that over 63.6% of

questions involved spatial thinking concepts. Most frequently asked questions focused on topics related to configuration, motion, and position (Kastens et al., 2011).

Spatial thinking within Earth Science spans from the atomic level, where the unique arrangement of atoms in a mineral affects properties such as cleavage or hardness, to the global level, with the movement of tectonic plates. Kastens and Ishikawa (2006) sort spatial thinking related tasks in Earth Science into three categories: 1) describing and interpreting objects, 2) comprehending spatial properties and processes, and 3) using space as a metaphor for other properties, such as time. To demonstrate proficiency in the first task, an individual must be able to clearly describe the shape of an object, classify an object by its shape, give meaning to the shape of a natural object, and recognize such shape in a distracting, noisy environment. To demonstrate proficiency in the second task, one must be able to remember the location and appearance of objects, describe the orientation and position of objects, construct and use maps, extrapolate from 1- and 2-D images to 3-D, and envision how objects can change shape or position. To demonstrate the third task, one should be able to use spatial thinking skills and strategies to think about non-spatial phenomena, such as temperature, pressure, and salinity (Kastens & Ishikawa, 2006).

Spatial ability tasks reveal one of the most consistent sex differences in cognitive abilities. Males consistently outperform females in large-scale and small-scale spatial ability tasks, such as mental rotation and spatial navigation (Toivainen et al., 2018; Tsigeman et al., 2023; Li et al., 2019; Voyer et al., 1995; Hromatko & Tadinac, 2006). Indeed, males outperform females on nearly every spatial thinking exercise, with the exception of object location memory tasks (De Goede & Postma, 2008). Hooven and her colleagues (2004) suggest that this gender imbalance occurs because individuals with higher testosterone levels typically show faster

response times and make fewer errors when completing mental rotation tasks (Hooven et al., 2004). Other studies suggest that the fact that males have a larger parietal lobe, a part of the brain responsible for processing sensory information, may be a factor (Wei et al., 2016; Hugdahl et al., 2006). Koscik, a graduate of the University of Iowa's Neuroscience Graduate Program hypothesized, "It's likely that the larger surface area in men's parietal lobes leads to an increase in functional columns, which are the processing unit in the cortex…This may represent a specialization for certain spatial abilities in men," (University of Iowa, 2008). Other researchers have concluded that the female supramarginal gyrus, a portion of the parietal lobe, and the parahippocampal gyrus, a portion of the medial temporal lobe, work less efficiently than in males (Li et al., 2019).

Studies regarding the cause for different brain structures are inconclusive, however. Some researchers are proponents of the Hunter-Gatherer theory, which hypothesizes that this difference is linked to evolution, as males predominantly took on the role of hunting (Sterling, 2014; Gurven & Hill, 2009). Hunting required strong spatial thinking skills to track prey and navigate landscapes (Silverman et al., 2007). As a result, males may have developed larger parietal lobes over time to adapt to these tasks and survive (Mithen, 1997). Additionally, this hypothesis purports that this evolutionary advantage most likely favored males with enhanced spatial abilities, which contributes to the sex differences in parietal lobe size and spatial skills between males and females today (Silverman et al., 2007; Koscik et al., 2009). Despite this theory's dominance in historical literature, modern research suggests that women were hunters as well, effectively disproving the Hunter-Gatherer theory (Anderson et al., 2023). Archaeological records have shown that women were buried with hunting spears and arrows, just like men were (Lacy & Ocobock, 2023). Additionally, exercise scientists have reported that women outperform

men in endurance tasks, such as marathons (Bam et al., 1997; Ocobock & Lacy, 2023). Both lines of evidence would support that women are as capable, if not better suited for, hunting, which heavily relies on endurance abilities (Liebenberg, 2006). Another study found that the majority of women hunt as well as forage in contemporary hunter-gatherer societies (Reyes‐García et al., 2020). Further research must be conducted on this topic, as it is an emerging field of interest that would drastically influence our understanding of human evolution and spatial thinking abilities.

In contrast to biological and evolutionary influences, other researchers seek to explain gender differences in spatial thinking abilities in a social context. Studies have shown that males and females experience unique social expectations and experiences (Parker et al., 2020). These expectations and experiences may impact their ability to develop spatial thinking skills (Clint et al., 2012; Baenninger & Newcombe, 1989). In other words, gender differences in spatial thinking may be caused by one's physical and social surroundings more so than by one's biological makeup. Lauer et al. (2019) performed a meta-analysis with statistics from over 30,000 children. They found no gender difference in mental rotation skills among preschool students, but a male advantage appeared to emerge as early as the age of six. Once small, this gender gap continued to grow through adolescence and adulthood (Lauer et al., 2019). Explanations as to why this occurred vary, but some posit that this difference may stem from the gendered nature of toys and sports that are marketed to young boys and girls (Raag, 1975; Moè et al., 2018; Tracy, 1987). Young boys are more likely to be pushed towards toys or opportunities that nurture their spatial thinking skills, such as building blocks, puzzles and playing sports, whereas young girls are more likely to be given toys or opportunities that nurture their social and verbal skills, such as dolls and roleplaying games (Blakemore & Centers, 2005; Moè et al., 2018; Trautner, 2016). Over

time, these habits and skills become ingrained and can influence spatial thinking abilities (Alexander, 2003). However, as societal attitudes toward gendered toys may be changing, it's uncertain how the habits and skills of the next generation will develop (MacDonald, 2023; Daly, 2016; Russell, 2022).

Anxiety is another possible factor that impacts female performance on assessments. Research has found that females are more likely than males to develop an anxiety disorder (Bahrami & Yousefi, 2011; McLean et al., 2011). Furthermore, females in STEM often experience heightened anxiety due to a variety of factors, such as lack of representation and support, discrimination, and pressure to perform in a male-dominated workforce (Stewart‐Williams & Halsey, 2021; Oliveira-Silva & De Lima, 2022). In turn, anxiety has been shown to negatively affect performance, which may contribute to gender differences in spatial ability-related test scores between females and males in STEM (Pelch, 2018).

Tsigeman and her colleagues (2023) found that gender differences in spatial thinking persist even with STEM experts. They found that while STEM experts displayed higher levels of spatial thinking than novices on average, continued practice with STEM tasks failed to close the gap between males and females. Further, the gap was even larger between males and females in the STEM expert group compared to the novices (Tsigeman et al., 2023).

Studies have found that spatial thinking abilities influence success in STEM, both in the short term (higher test scores) and long-term (ie: number of patents and publications), (Tsigeman et al., 2023; Gagnier et al., 2021; Shea et al., 2001; Gilbert, 2005). While interventions in the form of spatial thinking exercises have not proven effective in closing the gender gap, Sorby et al. (2013) proved that spatial thinking skills can improve with practice (Sorby et al., 2013). Researchers have found that regular interventions throughout the course of an academic semester

significantly improve students' spatial thinking skills, regardless of their proficiency prior to the interventions (Gold et al., 2018; Lowrie et al., 2018).

Spatial thinking skills are crucial for Earth scientists especially because Earth phenomena are described in terms of spatial concepts such as distance, scale, and size (Kastens et al., 2014; Kastens & Ishikawa, 2006). Without a deep understanding of these concepts, Earth scientists would fail to accurately understand the planet and, in turn, its complex processes. Since spatial thinking is often a challenging concept for humans to comprehend, further research into how individuals can best teach and learn spatial skills is crucial (Uttal et al., 2024).

My research aimed to investigate the potential impact of a semester-long class, focused on spatial thinking exercises, on students' spatial abilities. In Spring 2024, Vassar College offered a Structural Geology course taught by Professor Kirsten Menking. In this course, students learned to visualize in 3-dimensions, use geometric principles to predict what lies in the subsurface from surface observations, and conduct geologic mapping, among other skills. To develop or sharpen these skills, students engaged with a variety of spatial reasoning exercises such as creating cross-sections of different geologic structures from mapped surface features, predicting the outcrop pattern of rocks in undulating topography from a strike and dip measurement in a single location, and determining the strike and dip of bedding from apparent dips, among others. This classroom setting provided the opportunity to study spatial thinking skills in action. With this in mind, I sought to answer the following research questions: How, if at all, would a Structural Geology class impact students' spatial thinking abilities? Would there be observable gender disparities in spatial abilities? Would there be noticeable differences in the spatial thinking abilities of experts (Earth Science majors) versus novices (non-Earth Science majors)? In these assessments, would speed also be impacted by gender or expertise?

III. Methodology

In this study, my participants were 11 students at Vassar College enrolled in Structural Geology. Students ranged in age from 19 to 22. Six participants were women, three were men and two were non-binary. Two of the 11 students had never taken an Earth Science course before. The remaining participants had all taken at least two Earth Science courses. Students who had taken no Earth Science courses were classified as novices. Students who had taken more than one but fewer than six were considered intermediates. Students who had taken six or more Earth Science courses were classified as experts. I chose six courses as the benchmark because it's the minimum required for an Earth Science correlate. Participants were selected for this study because they were enrolled in the course. I had no control over the selection of students in the class, therefore, the size of my sample depended entirely on those who chose to enroll and were admitted into the class. Students were given the opportunity to opt out of this study, but all 11 consented to participate.

I employed a longitudinal observational research design with repeated measures to investigate how spatial thinking abilities among the group of students evolved throughout the first four weeks of the semester. I did not split participants into a control group and experimental group because the sample size was already small. I decided not to manipulate any variables because I was mainly focused on tracking the progression of spatial thinking skills throughout the semester. By refraining from manipulating variables such as practice opportunities or time, I was able to observe the natural development of spatial thinking abilities among the participants. I administered four spatial thinking tests throughout the first four weeks of the semester to track how students' spatial abilities changed. The repeated measures design allowed me to observe the changes in spatial thinking abilities over time within the same group of students.

Data collection spanned from January 22, 2024 - February 21, 2024. The data collection period began during the first lab of the semester and continued through the last lab before the midterm. I collected data through a pre-assessment survey and two post-assessment surveys, both in Google Forms, and four paper-and-pencil assessments. Before any data were collected, each participant signed a digital consent form that informed them of their rights as a participant (Appendix 1).

The pre-assessment survey (Appendix 2) was designed to gather demographic information on my participants. The survey was a Google Form with seven questions. These questions asked participants their age, gender identity, and experience with Earth Science classes, as well as the nature of games and toys they were exposed to, and gravitated towards, as children. The first post-assessment survey (Appendix 3) was a Google Form that consisted of five questions. It was administered after the first assessment. I asked participants how they felt about the assessment, what techniques they used to answer questions, and which questions they struggled with or found easy. The second post-assessment survey (Appendix 4) was administered after the last assessment by having students fill out a Google Form containing five questions that asked them to reflect on how their spatial thinking skills evolved throughout the semester.

In addition to obtaining informed consent, students were provided with the opportunity to maintain anonymity during their assessments and instead were identified by numbers. Potential risks to participants were minimized, as no identifiable personal information was utilized in my senior thesis. Furthermore, participants were free to withdraw from my study at any point, and their assessment scores did not impact their grades in Structural Geology.

The study has limitations that encompass various aspects of its design and measures. Firstly, it is important to note the small sample size, which may limit the generalizability of the

findings. Even though Vassar College is a smaller institution, the average class size is 17 students, and my study only had 11. Additionally, data collection was constrained to one month, preventing the tracking of students throughout the entire semester. In future research endeavors, I would recommend expanding the sample size to allow for a more comprehensive analysis. Further, achieving a more balanced distribution of gender representation would allow researchers to better explore gender differences within spatial thinking. Lastly, there is a need for standardized and accessible spatial tests that accurately measure individuals' spatial abilities because historical and current spatial tests are ridiculed for being inconsistent and unreliable (Hegarty & Waller, 2005; Thayaseelan et al., 2024). Uttal et al. (2024) corroborated these shortcomings, through surveys and interviews with fellow researchers of spatial thinking. They found that despite the existence of various spatial tests available online, they are often expensive and lack crucial information regarding their validity and reliability. For instance, historically, spatial thinking tests were developed using "WEIRD" samples (white/western, educated, industrialized, rich, and democratic) (Henrich et al., 2010). A lack of diversity may reflect a bias towards particular cultural, socioeconomic, and educational backgrounds. This bias limits the generalizability of the test results and may not accurately represent the spatial thinking abilities of individuals from diverse ethnic, socioeconomic, and cultural backgrounds (Bartlett & Camba, 2023). Additionally, some STEM-related spatial skills remain unmeasured by existing tests, such as non-rigid transformations (ie: bending, folding, or twisting) or the ability to externalize an internal representation (ie: creating a map) (Atit et al., 2020; Uttal et al., 2024).

Each assessment consisted of ten questions, eight multiple choice and two open-ended. Questions focused primarily on spatial visualization and reasoning tasks (See Appendix 5 for Assessment 1 questions and answers). Questions for each assessment were either pulled directly

from a variety of online resources or recreated in a similar fashion (Newton et al., n.d.; Ormand, n.d.; Hickson and Resnick, n.d.; *Free Spatial Reasoning Test Questions and Answers*, 2020). Each test included similar questions, with varying levels of difficulty. The topics covered in the four assessments were as follows: mental rotation, hole-punched paper, cube folding, cross-sections, slicing items, tangrams, viewing 3-D shapes from different angles, and maps.

Mental rotation problems required students to match a rotated object with its original position. These problems served to evaluate students' spatial visualization abilities and were crucial in examining gender disparities because this is a task where men typically outperform women (Li et al., 2019; Rahe et al., 2023). The hole-punched paper question involved spatial manipulation, reasoning and relations. It required students to understand the relationship between a drawing of initially folded paper, holes punched through it, and the positions of those holes once the paper was unfolded. The cube folding question prompted students to mentally rotate and transform a 2-D shape into a 3-D cube and determine the spatial relationships between the patterns on the faces of the unfolded cubes and how they would align when folded into a cube.

In questions about geologic cross-sections, students were asked to visualize how rock layers extend in three dimensions, even though they could only see a two-dimensional representation of the structure. They used their understanding of geological concepts, such as strike and dip, to interpret the orientation of the rock layers and infer how the layers continue beyond the visible portion. Questions about slicing items in different ways required students to mentally manipulate and visualize the object and the shapes resulting from the cuts. Tangram questions required students to manipulate geometric shapes and arrange them to form specific figures. Other questions required students to mentally visualize 3-D shapes from various

perspectives, while others tasked them with interpreting directional information and orientation depicted on a map.

Historically, assessments akin to the ones I used have been employed to evaluate spatial thinking. However, using spatial thinking assessments to evaluate student progress faces significant challenges. These challenges include a scarcity of reliable and valid spatial tests, disagreements about test constructs, inability to measure some STEM-relevant spatial skills, such as non-rigid transformations, and limited availability across age groups (Uttal et al., 2024; Brucato et al., 2022).

To create the four assessments I administered, I mostly used questions from existing spatial thinking assessments available online. Most of these assessments imposed time constraints on test-takers, however I did not. I chose not to impose a time limit on these assessments for a variety of reasons. Timed tests are often not entirely accurate representations of students' mastery of the content, rather how quickly and strategically they can answer questions (Ackerman & Ellingsen, 2016); Danthiir et al., 2005). Timed tests can also exclude students with disabilities who are legally required to have additional time on exams (Gernsbacher, 2015). The timed nature of tests often increases anxiety, negatively affecting test scores, and potentially exacerbating gender differences in spatial task performance, particularly for females (Tsigeman et al., 2023). To administer these assessments in the most inclusive and non-biased environment possible, students were given as much time as needed on the assessment. While time did not affect their score, it was still recorded to determine whether there was any correlation between time and gender, or time and expertise, and to determine whether students got faster at taking the tests throughout the semester. Each assessment was graded on a scale of 1-10. Partial credit was awarded for open-ended questions.

In this analysis, I sought to examine how students' spatial thinking abilities changed over time. I hypothesized that their abilities would improve over time, but not drastically, because this study only took place over the course of four weeks. I also sought to determine whether gender impacted students' scores. While previous research has shown that males consistently outperform females, I predicted that males would not outperform females in this case. At a historically women's liberal arts college, I suspected that there would be little to no significant difference in scores by gender because of Vassar's creation of a positive educational environment that seeks to promote female empowerment and academic success. Additionally, I wanted to test whether or not an individual's experience with spatial thinking influenced their score. I hypothesized that those deemed as experts, students who had taken more than six Earth Science classes, would outperform those deemed as novices, students who had not taken any Earth Science classes. Lastly, I was curious to see whether there would be any influence of gender or expertise on time to complete each assessment. I hypothesized that there would be no influence of gender on time to complete each assessment, but that experts would complete each assessment faster than novices or intermediates.

In order to analyze how student assessment scores changed over time, while considering the potential effects of gender and expertise, I decided to conduct several Analysis of Variance (ANOVA) tests. I chose ANOVA because it allowed me to compare averages across multiple groups, and in this case, examine variations in test scores at four different points in time.
I was able to incorporate time as a repeated measures factor, and gender and expertise as between-subject factors. Furthermore, ANOVA offered well-established methods for checking assumptions such as normality and homogeneity of variances, which ensured the reliability of results.

I conducted ANOVAs via JASP (M. Goss-Sampson, 2018), a statistical analysis software. For each student, I inputted the following data into JASP: gender, assigned level of expertise, time to complete each assessment, and test scores from each assessment (per question and per assessment). There were 11 types of problems incorporated into these assessments. Each type of problem was used at least twice throughout the assessment period. Two ANOVAs were run for each type of problem. For example, to analyze how gender impacted students' scores on tangram problems, the scores of every tangram problem, regardless of which assessment it came from, were analyzed with ANOVA separated by gender. Similarly, to analyze how expertise impacted students' scores on a particular type of problem (e.g., tangram), the scores of every problem, regardless of which assessment it came from, were analyzed with ANOVA separated by expertise.

With 11 types of questions and two ANOVAs (one for gender, one for expertise) for each type of question, 22 ANOVAs were run on the scores of the different types of questions. I also conducted a series of ANOVA tests to analyze changes in the time taken by students to complete assessments over the course of the study. Similarly, I analyzed overall scores using ANOVA, both collectively for all students across four assessments and separately based on gender and expertise. In total, I conducted 28 ANOVA tests (Appendix 6).

In order to run each test in JASP, I inputted the repeated measures factors (test sessions 1-4) and between subject factors (gender and expertise). Then, I asked the program to display descriptive statistics and estimates of effect size, specifically partial n^2 because it offers the most predictable and interpretable estimation of the effect size (National University, n.d.). Partial n^2 is calculated by dividing the sum of squares between groups by the total sum of squares. This value helped provide context for the p-value result that JASP produced. Within each test, a p-value

close to 0 indicates that the observed difference between groups (gender or expertise) was unlikely to be due to chance, whereas a p-value close to 1 suggested that there was no difference between the groups other than that due to chance. If a p-value was less than 0.05, I considered the relationship between test scores and either gender or expertise to be significant. Historically, 0.05 is a respectable threshold to differentiate significant results from non-significant results (Di Leo & Sardanelli, 2020).Whereas the p-value indicates whether differences between test scores, time to complete, or between-subject factors such as gender and expertise were statistically significant, the partial n^2 helps determine whether there is practical significance. In other words, this value sheds light on the strength of the significance. Partial n^2 also ranges from 0-1, with higher values indicating a stronger effect of the independent variable on the dependent variable. To interpret the partial n^2 value, I used the following guidelines, based on the National University Academic Success Center (ASC): a value of 0.01 indicates a small effect size, 0.06 indicates a medium effect size, and 0.14 or higher indicates a large effect size (National University, n.d.; Goss-Sampson, 2022).

After completing the ANOVAs, I next conducted an F-test to determine the significance of differences in means among the groups over time. Whereas the associated p-values indicate the probability of observing the data if there is truly no difference between the groups being compared, the F-test calculates the ratio of variability between group means to variability within groups. The F-test compares two types of variances: between-group variability and within-group variability. Between-group variability measures how much the means of different groups (e.g. male, female, non-binary) differ from each other. Within-group variability measures variability within each group. In other words, it measures how much individual scores deviate from their respective group means. The F-test determines whether differences observed between group

means are statistically significant compared to the variability within each group. If the ratio is large, it indicates that between-group variability is greater than within-group variability. This test assesses whether the differences observed between groups are larger than what would be expected due to random variation within groups. If the F-value exceeds the critical F value, it suggests significant differences between at least two of the groups (Blanca et al., 2023).

To find the critical F value, I first determined the degrees of freedom. In a repeated measures ANOVA, there are two types of degrees of freedom: degrees of freedom for the groups (df_between) and the degrees of freedom within groups (df_within). When comparing the effects of gender on performance, I had 3 groups: male, female, and non-binary. When comparing the effects of expertise on performance, I had 3 groups: novice, intermediates, and experts. To determine the degrees of freedom between groups I subtracted 1 from the number of groups, *k*. To determine the degrees of freedom within groups, I subtracted *k*, the number of groups, from *N,* the total number of observations. Next, I chose a significance level (alpha value), or p-value, of 0.05, since that is the most commonly used threshold for statistical significance. Then, I consulted an F distribution table for an alpha value of 0.05 and determined the critical F-value as the value at the intersection of the X column and Y row. If the calculated F-value from the analysis exceeded the F-critical value, the observed differences between group means were statistically significant at the chosen significance level.

Due to my study's small sample size and unequal group sizes (6 females, 3 males, 2 non-binary), it was likely that my data would violate sphericity. Sphericity assumes that the variance of differences between pairs of related groups (ie: Assessment 1 v. Assessment 2, Assessment 1 v. Assessment 3, etc.) in my data set are equal, which may not be the case. Therefore, assumption checks, including sphericity evaluation with Huynh-Feldt correction, were

conducted (Lund Research Ltd, 2018). I chose to run these checks because ensuring the validity of assumptions, such as sphericity, is important for accurately interpreting the results of repeated measures ANOVAs (Field, 2016). Additionally, making appropriate adjustments, such as applying the Huynh-Feldt correction if sphericity assumptions are violated, enhances the reliability of the statistical analysis (Lund Research Ltd, 2018). The Huynh-Feldt correction adjusts the degrees of freedom used in the F-test, which compares means in repeated measures ANOVA. The Hunyh-Feldt correction leads to more conservative p-values, which prevents inflated Type I Error Rates (described below) and reflects the correction for potential violations of the assumption of sphericity.

After running repeated measures ANOVAs and identifying significant differences between groups with reported p-values, I ran post hoc tests to further explore these differences and pinpoint where they existed. In other words, the p-value could indicate that there was a significant difference in test scores over time, but the post hoc test could pinpoint where (ie: between which assessments or which groups) these differences occurred. The post hoc tests required making multiple comparisons between groups.

When multiple comparisons are required, the risk of Type I errors increases because multiple tests increase the overall probability of incorrectly rejecting a true null hypothesis at least once, due to the random variability inherent in the data. A Type I error is known as a false positive. It occurs when the researcher falsely rejects a true null hypothesis, concluding that there is a significant difference or effect between groups when there is not. A null hypothesis, which is assumed to be true until there is significant evidence to reject it, suggests that there is no significant difference or relationship between variables.

To mitigate the increased risk of Type I errors associated with multiple comparisons, a Bonferroni correction was applied. Kastens used this correction in her work with repeated measures ANOVAs as well. The Bonferroni correction made the significance level stricter for each individual comparison. In the original repeated measures analyses, a significance level (alpha) of 0.05 was used. This means that there would be a 5% chance of incorrectly rejecting the null hypothesis (that there was no difference between groups defined by gender or expertise), a Type I error. However, the Bonferroni correction adjusted this significance level to 0.01, meaning that there is a 1% chance of incorrectly rejecting the null hypothesis when there is no true difference. In sum, this correction made it harder to declare a significant difference, decreasing the likelihood of declaring a significant difference between groups when there truly wasn't any. Lastly, I had JASP flag significant comparisons to identify which specific group differences were statistically significant.

IV. Results

After running a total of 28 ANOVAs, I primarily focused on interpreting the p-values from each test. I also compared the F-values to the critical F-value. The degrees of freedom between groups (df_between) was 2. The degrees of freedom within groups (df_within) was 8. According to the F-distribution table for an alpha of 0.05, with 2 as the degree of freedom for the numerator (df_between), and 8 as the degree of freedom for the denominator (df_within), the critical F-value was determined to be 4.459 (Purdue Department of Statistics, n.d.). If the F-value of a test is greater than the critical F-value, the null hypothesis is rejected, indicating that there is a significant difference between at least two of the groups.

The ANOVA that ran the overall scores of each student over the course of the study showed a p-value of 0.002 (Table 1). Since this p-value is less than 0.05, I rejected the null hypothesis and concluded that there was a significant difference in test scores over time. A partial n^2 value of 0.424 indicates a large effect size of the assessment number on test scores. Additionally, this test produced an F value of 6.635, which exceeds the critical F-value of this sample, 4.459. This indicates that the observed differences between group means are statistically significant at the chosen significance level of 0.05.

Table 1: Results of repeated measures ANOVA on the overall scores of four spatial thinking assessments given to eleven Structural Geology students

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η_p^2
Testing Sessions	Huynh-Feldt	11.849	2.813	4.212	6.635	0.002	0.424
Residuals	Huynh-Feldt	16.073	25.320	0.635			

Note. Type III Sum of Squares

However, post hoc comparisons show that there was only a significant difference in test scores between test 2, and every other test (Table 2). Overall, test scores did not change significantly between test 1 and test 4, with a reported p-value of 1.000. There was also not a significant difference in test scores from test 3 to test 4, with a p-value of 1.000. There was a significant difference in test scores between test 1 and 2, test 2 and 3, and test 2 and 4, with p-values of 0.009, 0.007 and 0.005, respectively. Figure 1 shows a visual representation of average test scores per testing session.

Table 2: Post-hoc comparisons of overall scores over time across four assessments taken by eleven students in Structural Geology

Post Hoc Comparisons - Testing Sessions

		Mean Difference	SE	t	p_{bonf}
A	B	−1.215	0.345	−3.521	0.009**
	C	0.035	0.345	0.101	1.000
	D	0.085	0.345	0.246	1.000
B	C	1.250	0.345	3.623	0.007**
	D	1.300	0.345	3.768	0.005**
C	D	0.050	0.345	0.145	1.000

** p < .01

Note. P-value adjusted for comparing a family of 6

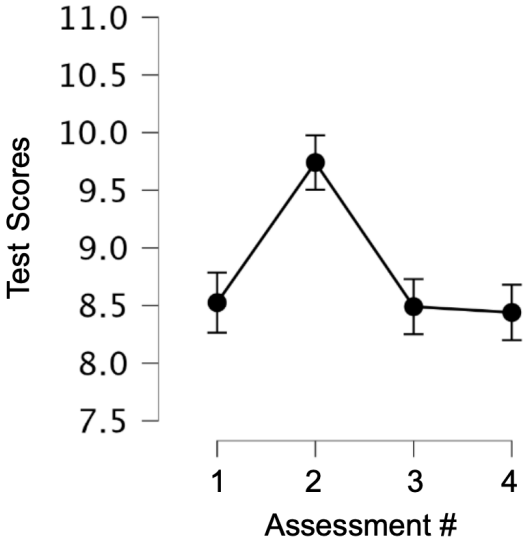


Figure 1: Structural Geology students' average overall test score for each assessment

To determine whether or not an individual's gender or expertise impacted their performance on spatial thinking activities, I ran ANOVAs for each type of problem assessed, with gender and expertise as between-subject factors. For example, when running ANOVAs on the tangram problem, JASP combined all the tangram scores of women, men, and non-binary individuals, regardless of assessment number. For the following question types, there was no

significant effect, as defined by a p-value of less than 0.05, of gender or expertise on student performance: tangram, hole punch, rotation, group rotation, cube, match cuts, draw cuts, 3D rotation (Appendix 6).

While gender did not impact an individual's score on map questions, expertise did. This ANOVA reported a p-value of 0.047, which is less than 0.05, indicating that there is a significant effect of expertise on performance on this type of question (Table 3). Additionally, an F-value of 4.900 exceeds the critical F-value of this sample, 4.459. This indicates that the observed differences between group means are statistically significant at the chosen significance level of 0.05. Indeed, the partial n^2 value of 0.583 indicates that there is a large effect of expertise on map questions. A post-hoc comparison did not show any significant difference between the mean scores of novices, intermediates and experts. However, a p-value of 0.055 indicates that the difference in scores between experts and intermediates was nearing significance (Table 4).

Table 3: Results of repeated measures ANOVAs on map questions, with expertise as the between subject factor

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.467	2	0.233	4.900	0.047	0.583
Residuals	0.333	7	0.048			

Note. Type III Sum of Squares

Table 4: Post-hoc comparisons of performance on map questions, broken down by expertise

Post Hoc Comparisons - Experience

		Mean Difference	SE	t	p_{bonf}
Expert	Novice	-2.082×10^{-16}	0.167	-1.249×10^{-15}	1.000
	Intermediate	0.333	0.109	3.055	0.055
Novice	Intermediate	0.333	0.178	1.871	0.311

* $p < .05$

Note. P-value adjusted for comparing a family of 3

Note. Results are averaged over the levels of: Map Testing Sessions

While expertise did not impact an individual's score on the block diagram questions, gender did. This ANOVA reported a p-value of 0.029, which is less than 0.05, indicating that there is a significant effect of gender on performance on this type of question (Table 5). Additionally, an F-value of 6.126 exceeds the critical F-value of this sample, 4.459. This indicates that the observed differences between group means are statistically significant at the chosen significance level of 0.05. Indeed, the partial n^2 value of 0.636 indicates that there is a large effect size of gender on performance for block diagram questions.

Table 5: Results of repeated measures ANOVAs on block diagram questions, with gender as the between subject factor

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.346	2	0.173	6.126	0.029	0.636
Residuals	0.198	7	0.028			

Note. Type III Sum of Squares

A post hoc comparison for the block diagram testing sessions, broken down by gender, showed that there was a significant difference between the scores of males and non-binary students, with a p-value of 0.031 (Table 6). There was not a significant difference between the scores of males and females, or between females and non-binary students. However, non-binary students outperformed male students on block diagram questions, with a mean difference in overall scores of 0.293.

Table 6: Post-hoc comparisons of performance on block diagram questions, broken down by gender

Post Hoc Comparisons - Gender

		Mean Difference	SE	t	p_{bonf}
(Non-binary)	Female	0.167	0.069	2.434	0.135
	Male	0.293	0.084	3.479	0.031*
Female	Male	0.125	0.069	1.827	0.331

Note. Results are averaged over the levels of: Block Diagram Testing Sessions
Note. P-value adjusted for comparing a family of 3
* p < .05

ANOVAs could not be run for the test scores on the type of question that asked individuals to identify a 3D object from above because there was no variance. In other words, every student, regardless of gender or expertise, received the same score every time for this problem type.

Neither gender nor expertise had a significant impact on the time it took students to complete the spatial thinking assessments. The p-value for time to complete assessments with respect to gender was 0.626 (Table 7). The F-value of 0.626 was below the critical F-value of this sample, 4.459. This indicates that the observed differences between group means were not statistically significant at the chosen significance level of 0.05. An effect size of 0.145 indicates that only approximately 14.5% of the variance in time to complete the test could be attributed to gender differences. If the sample size is small, which it is in this case, the statistical power of the ANOVA may not be enough to detect an effect even if it exists. Therefore, a large effect size points to a relationship between the independent and dependent variables, but not necessarily a statistically significant one. While the p-value indicates that the effect of gender on time to complete is not statistically significant, and the F-value is lower than the critical F-value, the partial n^2 value reveals a relationship between these variables. The partial n^2 value provides insight into the strength of the relationship between gender and time to complete. Despite a lack of statistical significance, as indicated by a p-value greater than 0.05 and the F-value being lower than the critical F-value, the partial n^2 value suggests that gender still plays a role in explaining variability in time to complete an assessment.

Table 7: Results of repeated measures ANOVAs on assessment completion time, with gender as the between-subject Factor

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	24.431	2	12.215	0.507	0.626	0.145
Residuals	144.625	6	24.104			

Note. Type III Sum of Squares

The p-value for time to complete assessments with respect to expertise was 0.791 (Table 8). This value indicates that expertise did not significantly impact the amount of time that students' took to complete an assessment. The F-value of 0.243 failed to exceed the critical F-value of 4.459, suggesting that the observed differences between group means were not statistically significant at the chosen significance level of 0.05. An effect size of 0.075 indicates that only approximately 7.5% of the variance in time to complete can be attributed to expertise differences. While the p-value shows that the effect of expertise is not statistically significant on time to complete, and the F-value is lower than the critical F-value, the partial n^2 value indicates a relationship between these variables.

Table 8: Results of repeated measures ANOVAs on block diagram questions, with expertise as the between subject factor

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	12.689	2	6.344	0.243	0.791	0.075
Residuals	156.367	6	26.061			

Note. Type III Sum of Squares

To complement the ANOVAs and post-hoc statistical analyses that determined whether there was a significant change in students' spatial thinking abilities, I administered a final survey to ask students whether they felt there was any change in their spatial thinking skills over the course of the study period. Of the eleven students involved in this study, nine reported that they

felt their spatial thinking skills had improved, one reported that they felt their spatial thinking skills had drastically improved, and one student felt that their spatial thinking skills stayed the same. When asked how challenging spatial thinking problems were now, compared to the beginning of the course, 18.2% of students responded with "somewhat," 45.5% responded with "a little," and 36.4% responded with "not at all." In survey responses after the first assessment, 66.7% of the students had reported these problems to be "a little challenging," and 33.3% reported that these problems were "somewhat challenging." Overall, it appears that student perceptions of the difficulty of spatial thinking problems shifted over the course of the semester even as their assessment scores showed little change, with a greater proportion initially finding them somewhat challenging, but later reporting them as only a little challenging or not at all.

V. Discussion

The results of my ANOVAs showed that there was no significant change in overall test scores between the first and last assessment. Students did not get significantly better or worse with spatial thinking problems. This result was contrary to my initial hypothesis that test scores would improve over time. One reason that no significant improvement occurred between the first and last assessment may be due to time. Tests 1 and 4 were administered only four weeks apart. This is a short time period, especially in comparison to the academic semester. It might be unrealistic to expect significant improvement only a quarter of the way through the spring semester. However, I was surprised to find that the only significant difference in overall test scores was in relation to test 2. If anything, I would have expected a significant change from the first assessment to the last assessment. The overall scores of test 2 were significantly different from those in 1, 3, and 4. While I expected a possible spike in test scores after the first

assessment, to account for students being more comfortable with the assessment format and routine, the interesting thing about the significant increase in scores for test 2 is that scores dropped back down for tests 3 and 4. Although it is unclear why there was a sudden increase in test scores for test 2, the time taken to complete each test did not influence the scores. Initially, I anticipated that as students became more familiar with spatial thinking topics, they would complete the tests more quickly, possibly due to improved understanding and problem-solving skills, or conversely, due to rushing and making more errors. However, there was no notable connection found between the time taken to complete the tests and the resulting scores.

Expertise did not have a significant effect on test scores or time to complete. While this result was surprising, I would argue that the liberal arts curriculum of Vassar College may have something to do with it. Since Vassar allows, and encourages, students to explore multiple academic disciplines, it is possible that even students who have not taken several Earth Science courses have still gained spatial thinking skills through other courses. At Vassar, a student majoring in English, for example, could easily take numerous classes about maps in Geography or visualization and perspective in an Art class. Another reason ANOVAs did not find a significant difference in test scores among those deemed novices, intermediates, and experts, could be due to my labeling of each group. I labeled students as one of the following strictly based on the number of Earth Science classes they had taken. As they were enrolled in an Earth Science course, and I assumed that Earth Science majors had more experience with spatial thinking due to the skills' prevalence and relevance in Earth Science courses, I associated Earth Science familiarity with spatial thinking abilities. I did not account for experiences outside of an Earth Science classroom, which created narrow, and likely inaccurate, definitions of novice, intermediate, and expert.

Results were likely also limited by the small sample size of this study, which included only 11 participants. Despite implementing the Bonferroni and Huynh-Feldt corrections to reduce the chance of error associated with small sample sizes, the non-significant p-values obtained suggest that results may still have been limited by the small sample size of the study. Therefore, further study with a larger sample size is recommended to enhance the robustness and generalizability of these findings.

When analyzing how students performed on a particular question, only one was significantly influenced by expertise. This question was one in which students had to interpret a map and follow a set directions to determine where they would end up. It is unclear why expertise only influenced scores on this kind of question. Similarly, only one type of problem was significantly influenced by gender. This question was the one in which students completed a partially filled-in block diagram, often in the form of a strike and dip diagram. It is unclear why gender only influenced scores on this kind of question, or why there was only a significant difference in performance between males and non-binary students.

Neither gender nor expertise significantly impacted the amount of time students took to complete each assessment. While I did not expect gender to have an influence on time, I did predict that experts would finish sooner than novices. I hypothesized that students who had been exposed to more spatial thinking exercises would be more familiar with the types of questions asked and the skills required to answer them, and would therefore complete the assessments more quickly. However, this was proven incorrect as there were no significant differences in time to complete among novices, intermediates, and experts.

The final survey administered to students gained information on their perceptions of how their spatial thinking skills evolved throughout the assessment period. It was not surprising that

the majority of students (91%) felt that their spatial thinking abilities had improved since the start of the semester, since Structural Geology was focused on fostering students' spatial thinking abilities.

VI. Conclusion

According to the ANOVAs, there was only one significant difference in scores over time, one significant difference in scores among genders for one type of question, and one significant difference in scores among expertise levels for one type of question. For the remaining 9 types of questions, there was no significant effect of gender or expertise. The small sample size of this study, as well as the unequal distribution of females, males and non-binary individuals, may have impacted the results. Unequal sample sizes result in unequal variances, which likely impacted the ANOVA results because the test assumes equal variances. Additionally, data collection took place over a short period of time, only about a quarter of the academic semester. This may explain why there was not a significant change in test scores from the first assessment to the last assessment. Additionally, the instructor for Structural Geology, Professor Kirsten Menking, felt that this group of students had unusually strong spatial thinking abilities to begin with based on 29 years of teaching experience.

My analysis of test scores over the first four weeks of Structural Geology reveals a lack of significant evolution in students' spatial thinking abilities. Additionally, expertise in Earth Science appeared to affect performance on only one of the 11 question types. Expertise also did not have any impact on time to complete. These findings suggest that prior Earth Science classes did not prove to be an advantage for these spatial thinking assessments. Gender appeared to affect scores on only one of the 11 question types, which suggests that gender did not

significantly impact overall test scores. I correctly hypothesized that gender would not significantly impact performance, due to the female empowerment associated with a historically women's liberal arts college, such as Vassar College. However, several students who did well on the spatial thinking assessments noted via the surveys I administered, that they had played Minecraft or similar video games in their adolescence. Milani and Di Blasio (2019) found that video games have the ability to enhance players' spatial thinking abilities, specifically mental rotation and spatial visualization. The immersive nature of gameplay experiences stimulate cognitive processes associated with spatial thinking (Milani & Di Blasio, 2019; Spence & Feng, 2010). These activities, which have gained popularity in recent years, may also influence spatial thinking abilities, and because of their accessibility, contribute to lessening the gender rift in modern teenagers. In fact, Feng et al. (2007), found that playing an action video game virtually eliminated the gender difference in spatial attention and decreased the gender disparity in mental rotation tasks.

Alongside these results, it is important to note the limitations of this study, including a small sample size and a narrow data collection window. Therefore, these findings should be considered preliminary, and future research efforts should aim to expand upon them to enhance both validity and reliability.

VII. Future Works

Were this study to be repeated, it would be helpful to perform further analyses on questions where there was a significant relationship between performance and gender or expertise. This would allow researchers to better understand what kind of individuals are performing better than others, and perhaps explore why. Additionally, it would be good to

increase the sample size of this study. A larger sample size would increase the statistical power, produce more generalizable findings, and reduce sampling error. While seeking to increase my sample size, I would also strive to recruit a balanced set of participants that represent a more equal distribution of genders and levels of expertise. If possible, I would track student progress throughout the entire semester, or year, to analyze how their spatial thinking abilities evolved beyond the span of four weeks. Additionally, the assigned labels of expertise could be reworked to account for other previous experiences.

In the initial pre-survey, I asked about students' history with activities related to spatial thinking, such as puzzles, video games, and Legos. However, I found it difficult to objectively categorize participants as novices, intermediates, or experts based solely on this information. To address this limitation in future surveys, I would advise researchers to gather additional details, including the frequency of participation in specific activities, perceived natural ability versus learned skill, exposure history, and other relevant factors. Alternatively, future studies could ask students to self-identify their level of expertise (ie: novice, intermediate, expert).

VIII. Appendix

This appendix includes surveys, sample questions and ANOVAs referenced in the text.

Appendix 1: Consent form that students read and signed before the beginning of data collection. Sent via Google Form.

Consent Form: Senior Thesis Participation

I have freely chosen to participate in this voluntary research study designed to provide information about the development of spatial reasoning skills through instruction in a Structural Geology course. The study will include four short spatial thinking assessments administered throughout the first half of the semester. I agree to permit the researcher, Monika Sweeney, to obtain, use and disclose the information provided as described below.

Conditions and Stipulations:

I understand that all information is confidential. I will not be personally identified in any reports. I agree to participate in this study for research purposes and that the data derived from this confidential survey may be made available to the general public in an anonymous fashion and in the form of Monika Sweeney's senior thesis presentation and library archive of her thesis.

I understand the study involves completing assessments and surveys about spatial reasoning and my previous exposure to childhood activities that can foster spatial reasoning skills.

I understand that my participation in this research survey is totally voluntary, and that declining to participate will involve no penalty in Kirsten's Structural Geology course. Choosing not to participate will not affect any beneficial opportunities, in any way. If I choose, I may withdraw my participation at any time. I also understand that if I choose to participate, that I may decline to answer any question that I am not comfortable answering.

By entering my name below, I freely provide consent and acknowledge my rights as a voluntary research participant as outlined above and provide consent to the researcher, Monika Sweeney, to use my information.

Monika Sweeney, student at Vassar College

msweeney@vassar.edu

If you wish to ask questions about your rights as a research participant or if you wish to voice any problems or concerns you may have about the study to someone other than the researcher, please reach out to my faculty sponsor Kirsten Menking, Professor of Earth Science and Director of Environmental Studies on the Althea Ward Clark Chair, at kimenking@vassar.edu.

If you consent, type your name below:

_____

Appendix 2: Pre-assessment survey that was administered to students during their first lab period of the semester

Spatial Thinking Assessment Pre-Survey

Please complete this brief survey BEFORE taking the spatial thinking assessment.

1. Please enter your name. If you prefer to remain anonymous, enter your textbook number. We will use this number to keep track of your spatial thinking progress and survey responses throughout the semester:
2. Please share your age:

3. Please share your gender identity:
4. Have you taken previous Earth Science classes? If so, list them below:
5. What kind of toys/games were you exposed to growing up? Select all that apply:
    a. Dolls
    b. Blocks/Legos
    c. Puzzles
    d. Sports equipment (basketball, tennis, soccer, etc.)
    e. Art materials (crayons, paint, pencils, etc.)
    f. Kitchen sets
    g. Cars
    h. Stuffed animals
    i. Memory games
    j. Magna-Tiles or PicassoTiles
    k. Origami
    l. Model-building kits (toy planes, cars, etc.)
    m. Minecraft, or similar video games that involve building
    n. Other:
6. What kinds of toys/games did you gravitate towards?
7. How do you feel about puzzles and brainteasers?

Appendix 3: Post-assessment survey that was administered to students during their first lab period of the semester, immediately after taking their first spatial thinking assessment

Spatial Thinking Assessment Pre-Survey

Please complete this brief survey AFTER taking the spatial thinking assessment.

1. Please enter your name below. If you prefer to remain anonymous, enter your textbook number. We will use this number to keep track of your spatial thinking progress and survey responses throughout the semester:
2. Did you find this activity challenging? (Select which most applies)
    a. Not at all
    b. A little
    c. Somewhat
    d. Very
    e. Extremely
3. How did you approach these questions? In other words, what techniques did you use to solve the problem?
    a. Hand gestures
    b. Drawing/Sketching/Writing it out
    c. Imagined it in my head
    d. Other:
4. Which type of problem(s) did you struggle with the most?
5. Which type of problem(s) did you find to be the easiest?

Appendix 4: Final survey that was administered to students immediately after taking their last spatial thinking assessment

Spatial Thinking Assessment Final Survey

Thank you so much for participating in my senior thesis! Please complete this final survey and reach out to msweeney@vassar.edu with any questions

1. Compared to the beginning of this course, I feel that my spatial thinking skills have....
    a. Drastically worsened
    b. Worsened
    c. Stayed the same
    d. Improved
    e. Drastically improved
    f. Other:
2. What activities did you find most helpful/beneficial for the development of your spatial thinking skills?
    a. Lectures
    b. Labs (Hands-on activities)
    c. Problem sets
    d. Exams
    e. Other:
3. Think back to the assessments I administered in the first half of this course (ie: cube folding, map navigation, mental rotation, etc.). Now that you've taken four of these assessments, how challenging do you find these kinds of activities?
    a. Not at all
    b. A little
    c. Somewhat

d. Very

e. Extremely

4. Anything else you'd like to add?:

Appendix 5: Spatial thinking assessment 1, with answer key.

Name:

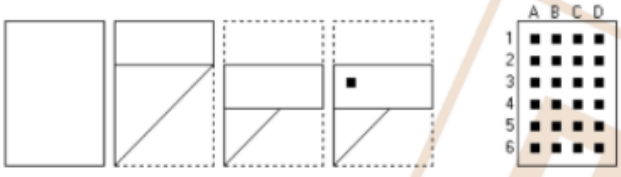
Directions: Answer each question to the best of your ability. This is **NOT** a test. You will not be graded on this, nor will it affect your grade in this course. It will only be used to track the evolution of your spatial thinking skills throughout the semester, research which will be used for Monika's senior thesis project. *Note: If you spend more than 10 minutes on any problem, just move on to the next.*

Start Time: _____ **End Time:** _____

Assessment 1

1. Go to https://mathigon.org/tangram. Use the shapes provided to create the **teapot**. Keep you screen open until Monika has come around to check it. **(Tangram)**

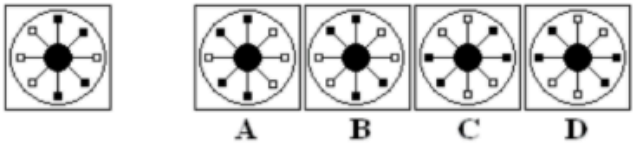
2. The drawings show a piece of paper that has been folded. The dashed lines indicate the original sheet; each drawing represents a single fold. The black squares show where a hole was punched. Where do the holes appear when the sheet is unfolded? **(Hole Punch)**



A	B	C	D
3A,2A,6D	3A,5A,6D	3A,5A,3D	3A,2A

A B C D

3. Which figure is identical to the first? The answer may be rotated. **(Rotation)**

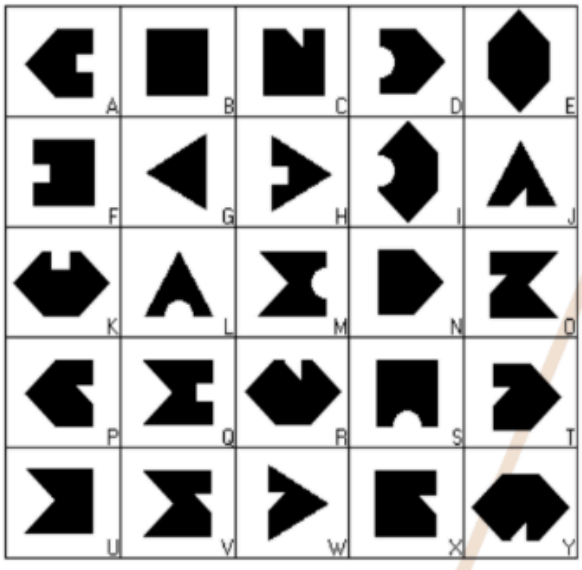


A B C D

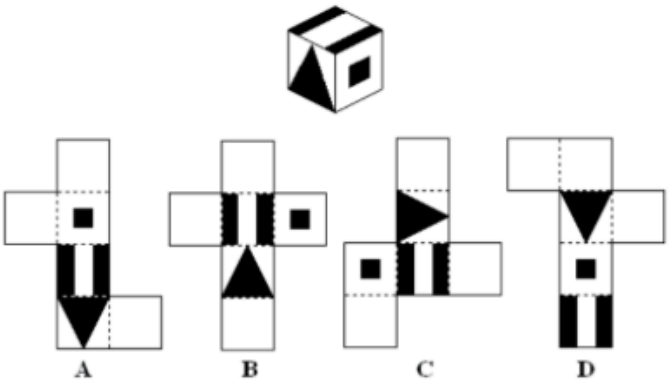
4. All of the shapes in Group 1 appear in Group 2, although some of them may be rotated. Watch out for the ones that are reflected! Which shapes in Group 2 correspond to the shapes in Group 1 (1-5)? **(Group Rotation)**

Group 1

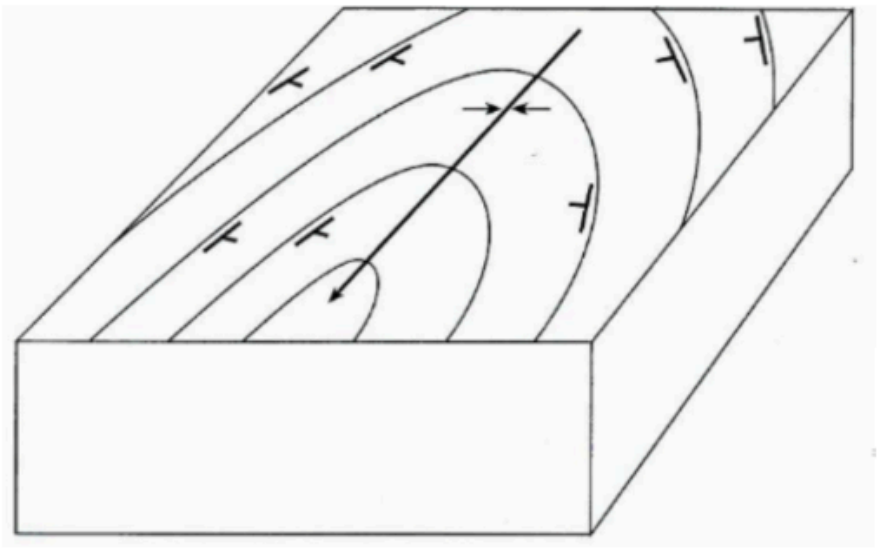
Group 2



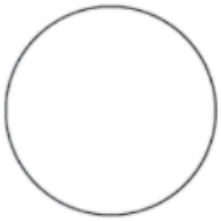
5. Which pattern can be folded to make the cube shown? **(Cube)**

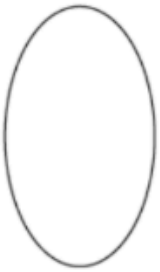


6. Complete the blank sides of the diagram below. The t-shaped symbol is a strike and dip symbol. The vertical leg of the T is pointing in the direction that the rock layers are dipping. **(Block Diagram)**



7. An ice cream cone is shown below. It is cut in three spots (A, B and C). Match the cuts to the corresponding shapes they would create. **(Match Cuts)**

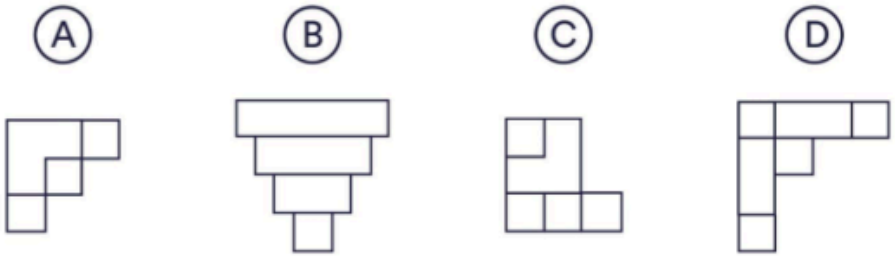




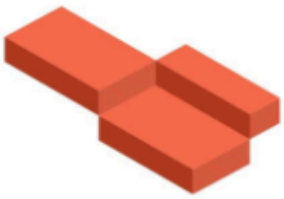
_____ _____ _____

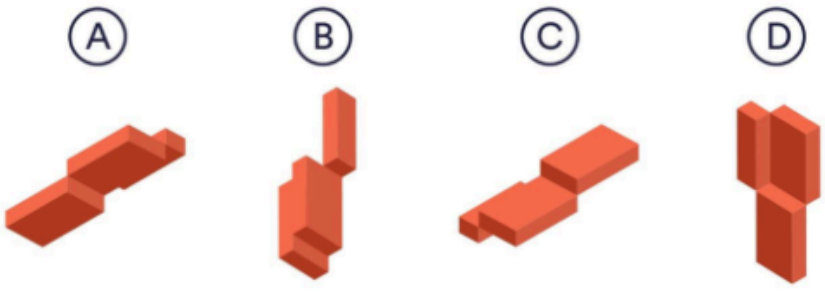
8. The ice cream is cut in another spot (D). Draw the corresponding shape that the slice would produce. **(Draw Cuts)**

9. What would this 3D shape look like from above? **(3-D Above)**



10. Which of the given shapes is the same 3D shape below but in a different position? **(3-D Rotated)**

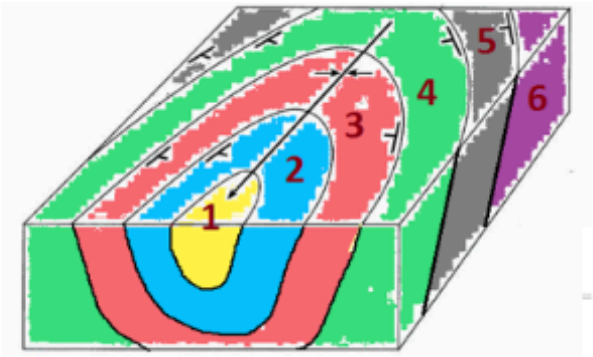


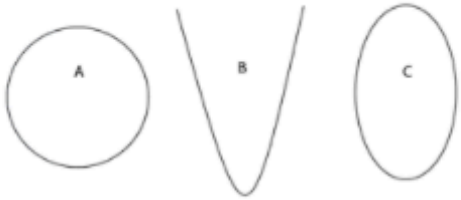


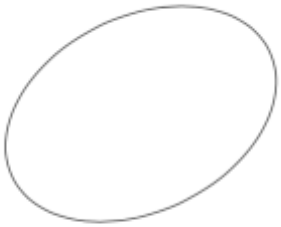
Assessment 1: Answer Key

1. 

2. A
3. A
4. 1(N), 2(Q), 3(T), 4(F), 5(X)
5. B

6. 

7. 

8. 

9. A
10. D

Appendix 6: Results from repeated measures ANOVAs on the following problems, with respect to gender and expertise: tangram (6-A), hole punch (6-B), rotation (6-C), group rotation (6-D), cube (6-E), match cuts (6-F), draw cuts (6-G), 3-D rotation (6-H); Results from repeated measures ANOVAs, with respect to expertise: block diagram (6-I); Results from repeated measures ANOVAs, with respect to gender: maps (6-J); Results from repeated measures ANOVAs on time to complete over time, with respect to gender (6-K), and time to complete over time, with respect to expertise (6-K); Results from repeated measures ANOVAs on overall scores over time, with respect to gender (6-L), and overall scores over time, with respect to expertise (6-L).

6-A: Tangram scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.017	2	0.008	0.280	0.764	0.074
Residuals	0.208	7	0.030			

Note. Type III Sum of Squares

6-A: Tangram scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.017	2	0.008	0.280	0.764	0.074
Residuals	0.208	7	0.030			

Note. Type III Sum of Squares

6-B: Hole punch scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.242	2	0.121	1.455	0.289	0.267
Residuals	0.667	8	0.083			

Note. Type III Sum of Squares

6-B: Hole punch scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.409	2	0.205	3.273	0.092	0.450
Residuals	0.500	8	0.062			

Note. Type III Sum of Squares

6-C: Rotation scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.100	2	0.050	0.700	0.528	0.167
Residuals	0.500	7	0.071			

Note. Type III Sum of Squares

6-C: Rotation scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.058	2	0.029	0.377	0.699	0.097
Residuals	0.542	7	0.077			

Note. Type III Sum of Squares

6-D: Group rotation scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.100	2	0.050	0.700	0.528	0.167
Residuals	0.500	7	0.071			

Note. Type III Sum of Squares

6-D: Group rotation scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.058	2	0.029	0.377	0.699	0.097
Residuals	0.542	7	0.077			

Note. Type III Sum of Squares

6-E: Cube scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.222	2	0.111	0.298	0.751	0.078
Residuals	2.611	7	0.373			

Note. Type III Sum of Squares

6-E: Cube scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.167	2	0.083	0.219	0.809	0.059
Residuals	2.667	7	0.381			

Note. Type III Sum of Squares

6-F: Match cuts scores over time, with respect to gender

Between Subjects Effects						
Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.028	2	0.014	0.198	0.825	0.054
Residuals	0.488	7	0.070			

Note. Type III Sum of Squares

6-F: Match cuts scores over time, with respect to expertise

Between Subjects Effects						
Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.047	2	0.024	0.353	0.715	0.092
Residuals	0.469	7	0.067			

Note. Type III Sum of Squares

6-G: Draw cuts scores over time, with respect to gender

Between Subjects Effects						
Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.150	2	0.075	0.600	0.575	0.146
Residuals	0.875	7	0.125			

Note. Type III Sum of Squares

6-G: Draw cuts scores over time, with respect to expertise

Between Subjects Effects						
Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.140	2	0.070	0.552	0.599	0.136
Residuals	0.885	7	0.126			

Note. Type III Sum of Squares

6-H: 3D rotation scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.100	2	0.050	0.350	0.716	0.091
Residuals	1.000	7	0.143			

Note. Type III Sum of Squares

6-H: 3D rotation scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.058	2	0.029	0.196	0.826	0.053
Residuals	1.042	7	0.149			

Note. Type III Sum of Squares

6-I: Block diagram scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	0.106	2	0.053	0.847	0.469	0.195
Residuals	0.438	7	0.063			

Note. Type III Sum of Squares

6-J: Maps scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	0.133	2	0.067	0.700	0.528	0.167
Residuals	0.667	7	0.095			

Note. Type III Sum of Squares

6-K: Time to complete over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Gender	24.431	2	12.215	0.507	0.626	0.145
Residuals	144.625	6	24.104			

Note. Type III Sum of Squares

6-K: Time to complete over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Experience	12.689	2	6.344	0.243	0.791	0.075
Residuals	156.367	6	26.061			

Note. Type III Sum of Squares

6-L: Overall scores over time, with respect to gender

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
ATime (Mins)	1.697	1	1.697	0.536	0.540	0.211
BTime (Mins)	2.286	1	2.286	0.722	0.485	0.265
CTime (Mins)	4.022	1	4.022	1.270	0.377	0.388
DTime (Mins)	0.063	1	0.063	0.020	0.901	0.010
Gender	0.683	2	0.342	0.108	0.903	0.097
Residuals	6.333	2	3.166			

Note. Type III Sum of Squares

6-L: Overall scores over time, with respect to expertise

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
ATime (Mins)	2.562	1	2.562	1.014	0.420	0.336
BTime (Mins)	2.184	1	2.184	0.864	0.451	0.302
CTime (Mins)	4.063	1	4.063	1.608	0.332	0.446
DTime (Mins)	0.503	1	0.503	0.199	0.699	0.091
Experience	1.963	2	0.982	0.389	0.720	0.280
Residuals	5.053	2	2.526			

Note. Type III Sum of Squares

References

Ackerman, P. L., & Ellingsen, V. J. (2016). Speed and accuracy indicators of test performance under different instructional conditions: Intelligence correlates. *Intelligence*, *56*, 1–9. https://doi.org/10.1016/j.intell.2016.02.004

Alexander, G. M. (2003). An evolutionary perspective of sex-typed toy preferences: pink, blue, and the brain. *Archives of Sexual Behavior*, *32*(1), 7–14. https://doi.org/10.1023/a:1021833110722

Anderson, A., Chilczuk, S., Nelson, K. D., Ruther, R., & Wall-Scheffler, C. M. (2023). The Myth of Man the Hunter: Women's contribution to the hunt across ethnographic contexts. *PloS One*, *18*(6), e0287101. https://doi.org/10.1371/journal.pone.0287101

Atit, K., Uttal, D. H., & Stieff, M. (2020). Situating space: using a discipline-focused lens to examine spatial thinking skills. *Cognitive Research*, *5*(1). https://doi.org/10.1186/s41235-020-00210-z

Baenninger, M., & Newcombe, N. S. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles*, *20*(5–6), 327–344. https://doi.org/10.1007/bf00287729

Bahrami, F., & Yousefi, N. (2011). Females are more anxious than males: a metacognitive perspective. *Iranian journal of psychiatry and behavioral sciences*, 5(2), 83–90. https://pubmed.ncbi.nlm.nih.gov/24644451

Bam, J., Noakes, T. D., Juritz, J., & Dennis, S. C. (1997). Could women outrun men in ultramarathon races? *Medicine and Science in Sports and Exercise*, *29*(2), 244–247. https://doi.org/10.1097/00005768-199702000-00013

Mithen, S. (1997). [Review of *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, by J. H. Barkow, L. Cosmides, & J. Tooby]. *Journal of Anthropological Research*, *53*(1), 100–102. http://www.jstor.org/stable/3631124

Bartlett, K. A., & Camba, J. D. (2023). Gender Differences in Spatial Ability: a Critical Review. *Educational Psychology Review*, *35*(1). https://doi.org/10.1007/s10648-023-09728-2

Bednarz, R. S., & Lee, J. (2011). The components of spatial thinking: empirical evidence. *Procedia: Social & Behavioral Sciences*, *21*, 103–107. https://doi.org/10.1016/j.sbspro.2011.07.048

Blakemore, J. E., & Centers, R. E. (2005). Characteristics of boys' and girls' toys. *Sex Roles*, *53*(9–10), 619–633. https://doi.org/10.1007/s11199-005-7729-0

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023). Repeated measures ANOVA and adjusted F-tests when sphericity is violated: which procedure is best? *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1192453

Brucato, M., Frick, A., Pichelmann, S., Nazareth, A., & Newcombe, N. S. (2022). Measuring Spatial Perspective Taking: Analysis of four measures using item response theory. Topics in Cognitive Science, 15(1), 46–74. https://doi.org/10.1111/tops.12597

Chow, C., Epp, J. R., Lieblich, S. E., Barha, C. K., & Galea, L. A. (2013). Sex differences in neurogenesis and activation of new neurons in response to spatial learning and memory. *Psychoneuroendocrinology*, *38*(8), 1236–1250. https://doi.org/10.1016/j.psyneuen.2012.11.007

Clint, E. K., Sober, E., Garland, T., & Rhodes, J. S. (2012). Male superiority in spatial navigation: adaptation or side effect? *The Quarterly Review of Biology*, *87*(4), 289–313. https://doi.org/10.1086/668168

Sterling, K. (2014). Man the Hunter, Woman the Gatherer? The Impact of Gender Studies on Hunter-Gatherer Research (A Retrospective). In *Oxford University Press eBooks*. https://doi.org/10.1093/oxfordhb/9780199551224.013.032

Daly, N. (2017). How today's toys may be harming your daughter. *National Geographic*. https://www.nationalgeographic.com/magazine/article/gender-toys-departments-piece

Danthiir, V., Wilhelm, O., & Schacht, A. (2005). Decision speed in intelligence tasks: correctly an ability? *Psychology Science*, *47*(2), 200–229. De Goede, M., & Postma, A. (2008). Gender differences in memory for objects and their locations: A study on automatic versus controlled encoding and retrieval contexts. *Brain and Cognition*, *66*(3), 232–242. https://doi.org/10.1016/j.bandc.2007.08.004

Di Leo, G., & Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, *4*(1), 18. https://doi.org/10.1186/s41747-020-0145-y

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*(10), 850–855. https://doi.org/10.1111/j.1467-9280.2007.01990.x

Field, A. (2016). *Repeated Measures ANOVA* (pp. 1–3). Retrieved January 2, 2024, from https://discoveringstatistics.com/docs/repeatedmeasures.pdf

Free spatial reasoning test questions and answers. (2020, May 11). Practice Aptitude Tests. Retrieved January 2, 2024, from https://www.practiceaptitudetests.com/free-spatial-reasoning-test-questions-and-answers/

Gagnier, K. M., Holochwost, S. J., & Fisher, K. (2021). Spatial thinking in science, technology, engineering, and mathematics: Elementary teachers' beliefs, perceptions, and self‐efficacy. *Journal of Research in Science Teaching (Print)*, *59*(1), 95–126. https://doi.org/10.1002/tea.21722

Gauvain, M. (1992). Sociocultural influences on the development of spatial thinking. *Children's Environments*, *9*(1), 27–36. https://www.jstor.org/stable/41514848

Gernsbacher M. A. (2015). Diverse Brains. *The general psychologist*, *49*(2), 29–37. https://pubmed.ncbi.nlm.nih.gov/28090598

Gilbert, J.K. (2005). Visualization: A Metacognitive Skill in Science and Science Education. In: Gilbert, J.K. (eds) Visualization in Science Education. Models and Modeling in Science Education, vol 1. Springer, Dordrecht. https://doi.org/10.1007/1-4020-3613-2_2

Gold, A., Pendergast, P. M., Ormand, C. J., Budd, D. A., & Mueller, K. (2018). Improving spatial thinking skills among undergraduate geology students through short online training exercises. *International Journal of Science Education*, *40*(18), 2205–2225. https://doi.org/10.1080/09500693.2018.1525621

Goss-Sampson, M. (2018). *JASP 0.9.* https://static.jasp-stats.org/Statistical%20Analysis%20in%20JASP%20-%20A%20Students%20Guide%20v1.0.pdf

Goss-Sampson, M. (2022). Statistical Analysis in *JASP: A Guide for Students, 5th Edition* https://jasp-stats.org/wp-content/uploads/2022/04/Statistical-Analysis-in-JASP-A-Students-Guide-v16.pdf

Gurven, M., & Hill, K. (2009). Why do men hunt? *Current Anthropology*, *50*(1), 51–74. https://doi.org/10.1086/595620

Hegarty, M. (2010). Components of spatial intelligence. *Psychology of learning and motivation, 52,* 265–297. https://doi.org/10.1016/s0079-7421(10)52007-3

Hegarty, M., & Waller, D. A. (2005). Individual Differences in Spatial Abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 121–169). chapter, Cambridge: Cambridge University Press.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Hickson, T. and Resnick, I. (n.d.). *Sketching block diagrams*. Teaching Activities. https://serc.carleton.edu/spatialworkbook/activities/sketching_blocks.html

Hooven, C., Chabris, C., Ellison, P., & Kosslyn, S. (2004). The relationship of male testosterone to components of mental rotation. *Neuropsychologia*, *42*(6), 782–790. https://doi.org/10.1016/j.neuropsychologia.2003.11.012

Hromatko, I., & Tadinac, M. (2006). Testosterone levels influence spatial ability: Further evidence for curvilinear relationship. Review of Psychology, 13(1), 27–34. https://bib.irb.hr/datoteka/267783.review2006.pdf

Hugdahl, K., Thomsen, T., & Ersland, L. (2006). Sex differences in visuo-spatial processing: An fMRI study of mental rotation. *Neuropsychologia*, *44*(9), 1575–1583. https://doi.org/10.1016/j.neuropsychologia.2006.01.026

Ishikawa, T., & Newcombe, N. S. (2021). Why spatial is special in education, learning, and everyday activities. *Cognitive Research*, *6*(1). https://doi.org/10.1186/s41235-021-00274-5

Kastens, K. (2021). Cultivating a new field at the boundary between geoscience and education research. *Perspectives of Earth and Space Scientists*, *2*(1). https://doi.org/10.1029/2021cn000149

Kastens, K. A., & Ishikawa, T. (2006). Spatial thinking in the geosciences and cognitive sciences: A cross-disciplinary look at the intersection of the two fields. In C. A. Manduca & D. W. Mogk (Eds.), *Earth and Mind: How Geologists Think and Learn about the Earth*. Boulder, CO: Geological Society of America. https://doi.org/10.1130/2006.2413(05)

Kastens, K., & Passow, M. (2012). Opening a conversation about spatial thinking in earth science. *National Earth Science Teachers Association*, *XXVIII*(4), 37–40. https://earth2class.org/site/wp-content/uploads/2015/06/OpeningSpatialThinkingConv.pdf

Kastens, K., Pistolesi, L., & Passow, M. J. (2014). Analysis of spatial concepts, spatial skills and spatial representations in New York State Regents Earth Science examinations. *Journal of Geoscience Education*, *62*(2), 278–289. https://doi.org/10.5408/13-104.1

Kastens, K., Pistolesi, L., Passow, M., & Lamont-Doherty Earth Observatory. (2011, October 9). *Spatial thinking in the New York State High School Earth Science Exam* (By Geological Society of America). Geological Society of America Conference. https://oceansofdata.org/sites/oceansofdata.org/files/Spatial_Regents_GSA2011_v3.pdf

King, M. G., Katz, D. P., Thompson, L. A., & Macnamara, B. N. (2019). Genetic and environmental influences on spatial reasoning: A meta-analysis of twin studies. *Intelligence*, *73*, 65–77. https://doi.org/10.1016/j.intell.2019.01.001

Koscik, T. R., O'Leary, D., Moser, D. J., Andreasen, N. C., & Nopoulos, P. (2009). Sex differences in parietal lobe morphology: Relationship to mental rotation performance. *Brain and Cognition*, *69*(3), 451–459. https://doi.org/10.1016/j.bandc.2008.09.004

Lacy, S., & Ocobock, C. (2023). Woman the hunter: The archaeological evidence. *American Anthropologist*, *126*(1), 19–31. https://doi.org/10.1111/aman.13914

Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin*, *145*(6), 537–565. https://doi.org/10.1037/bul0000191

Li, Y., Kong, F., Luo, Y., Zeng, S., Lan, J., & You, X. (2019). Gender Differences in Large-Scale and Small-Scale Spatial Ability: A Systematic review based on Behavioral and Neuroimaging research. *Frontiers in Behavioral Neuroscience*, *13*. https://doi.org/10.3389/fnbeh.2019.00128

National University. (n.d.). *LibGuides: Statistics Resources: Partial ETA squared*. https://resources.nu.edu/statsresources/eta

https://resources.nu.edu/statsresources/eta

Liebenberg, L. (2006). Persistence hunting by modern Hunter‐Gatherers. *Current Anthropology*, *47*(6), 1017–1026. https://doi.org/10.1086/508695

Lowrie, T., Logan, T., Harris, D., & Hegarty, M. (2018). The impact of an intervention program on students' spatial reasoning: student engagement through mathematics-enhanced learning activities. *Cognitive Research*, *3*(1). https://doi.org/10.1186/s41235-018-0147-y

MacDonald, M. (2023, January 9). Stop gendering toys! *Medium*. https://medium.com/@matthew_macdonald/stop-gendering-toys-715844e577ed

McLean, C. P., Asnaani, A., Litz, B. T., & Hofmann, S. (2011). Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *Journal of Psychiatric Research*, *45*(8), 1027–1035. https://doi.org/10.1016/j.jpsychires.2011.03.006

McNeal, P., & Petcovic, H. L. (2020). Spatial thinking and fluid Earth science education research. *Journal of Geoscience Education*, *68*(4), 289–301. https://doi.org/10.1080/10899995.2020.1768007

Milani, L., & Di Blasio, P. (2019). Positive effects of videogame use on visuospatial competencies: The impact of visualization style in preadolescents and adolescents. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01226

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

Newton, P., Bristoll, H., & Psychometric Success. (n.d.). *SPATIAL ABILITY— PRACTICE TEST*. Retrieved January 21, 2024, from https://psychometric-success.com/test-pdfs/PsychometricSuccessSpatialAbility-PracticeTest1-8ede.pdf

Ocobock, C., & Lacy, S. (2023). Woman the hunter: The physiological evidence. *American Anthropologist*, *126*(1), 7–18. https://doi.org/10.1111/aman.13915

Oliveira-Silva, L. C., & De Lima, M. C. C. (2022). Mental health of women in STEM. *Psico*, *53*(1), e38473. https://doi.org/10.15448/1980-8623.2022.1.38473

Ormand, C. (n.d.). *Slices through 3D objects*. Spatial Thinking Workbook. https://serc.carleton.edu/spatialworkbook/activities/3D_slices.html

Parker, K., Horowitz, J., & Stepler, R. (2020, August 6). *2. Americans see different expectations for men and women | Pew Research Center*. Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/2017/12/05/americans-see-different-expectations-for-men-and-women/

Pelch, M. A. (2018). Gendered differences in academic emotions and their implications for student success in STEM. *International Journal of STEM Education*, *5*(1). https://doi.org/10.1186/s40594-018-0130-7

Purdue Department of Statistics. (n.d.). *Critical values of the F -Distribution: A = 0.05*. Purdue University. https://www.stat.purdue.edu/~lfindsen/stat503/F_alpha_05.pdf

Raag, T. (1975). Influences of Social Expectations of Gender, Gender Stereotypes, and Situational Constraints on Children's Toy Choices. *Sex Roles*, *41*(11).

Rahe, M., Schürmann, L., & Jansen, P. (2023). Self-concept explains gender differences in mental rotation performance after stereotype activation. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1168267

Rebelsky, F. (1964). Adult perception of the horizontal. *Perceptual and Motor Skills*, *19*(2), 371–374. https://doi.org/10.2466/pms.1964.19.2.371

Reyes‐García, V., Díaz-Reviriego, I., Duda, R., Fernández‐Llamazares, Á., & Gallois, S. (2020). "Hunting otherwise." *Human Nature (Hawthorne, N.Y.)*, *31*(3), 203–221. https://doi.org/10.1007/s12110-020-09375-4

Rochberg, F. (2002). A consideration of Babylonian astronomy within the historiography of science. *Studies in History and Philosophy of Science*, *33*(4), 661–684. https://doi.org/10.1016/s0039-3681(02)00022-5

Russell, H. (2022, February 18). Lego to remove gender bias from its toys after findings of child survey. *The Guardian*. https://www.theguardian.com/lifeandstyle/2021/oct/11/lego-to-remove-gender-bias-after-survey-shows-impact-on-children-stereotypes

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

Silverman, I., Choi, J., & Peters, M. (2007). The Hunter-Gatherer Theory of Sex Differences in Spatial Abilities: Data from 40 Countries. *Archives of Sexual Behavior*, *36*(2), 261–268. https://doi.org/10.1007/s10508-006-9168-6

Sorby, S. A. (2007). Developing 3D spatial skills for engineering students. *Australasian Journal of Engineering Education*, *13*(1), 1–11. https://doi.org/10.1080/22054952.2007.11463998

Sorby, S. A., Casey, B. M., Veurink, N., & Dulaney, A. (2013). The role of spatial training in improving spatial and calculus performance in engineering students. *Learning and Individual Differences*, *26*, 20–29. https://doi.org/10.1016/j.lindif.2013.03.010

Spence, I., & Feng, J. (2010). Video games and spatial cognition. *Review of General Psychology*, *14*(2), 92–104. https://doi.org/10.1037/a0019491

Lund Research Ltd. (2018). *Sphericity*. Laerd Statistics. https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide.php

Steele, J. M., & Ganguli, S. (2022). Babylonian records of transient astronomical phenomena. *Astronomische Nachrichten*, *343*(6–7). https://doi.org/10.1002/asna.20220031

Stewart‐Williams, S., & Halsey, L. G. (2021). Men, women and STEM: Why the differences and what should be done? *European Journal of Personality (Print)*, *35*(1), 3–39. https://doi.org/10.1177/0890207020962326

Thayaseelan, K., Zhai, Y., Li, S., & Liu, X. (2024). Revalidating a measurement instrument of spatial thinking ability for junior and high school students. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1). https://doi.org/10.1186/s43031-024-00095-8

Toivainen, T., Pannini, G., Papageorgiou, K. A., Malanchini, M., Rimfeld, K., Shakeshaft, N. G., & Kovas, Y. (2018). Prenatal testosterone does not explain sex differences in spatial ability. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-31704-y

Tracy, D. M. (1987). Toys, spatial ability, and science and mathematics achievement: Are they related? *Sex Roles*, *17*(3–4), 115–138. https://doi.org/10.1007/bf00287620

Trautner, T. (2016, November 30). *Dangers of gender-based toys*. Michigan State University. Retrieved February 11, 2024, from https://www.canr.msu.edu/news/dangers_of_gender_based_toys#:~:text=Play%20with%20masculine%20toys%20is,many%20diverse%20interests%20and%20skills

Tsigeman, E., Likhanov, M., Budakova, A. V., Akmalov, A. F., Sabitov, I., Alenina, E., Bartseva, K., & Kovas, Y. (2023). Persistent gender differences in spatial ability, even in STEM experts. *Heliyon*, *9*(4), e15247. https://doi.org/10.1016/j.heliyon.2023.e15247

University of Iowa. (2008). Sex difference on spatial skill test linked to brain structure. *Brain and Cognition*. Retrieved March 10, 2024, from https://www.eurekalert.org/news-releases/742470

Uttal, D. H., McKee, K., Simms, N., Hegarty, M., & Newcombe, N. S. (2024). How can we best assess spatial skills? practical and conceptual challenges. *Journal of Intelligence*, *12*(1), 8. https://doi.org/10.3390/jintelligence12010008

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Uttal, D. H., Miller, D. I., & Newcombe, N. S. (2013). Exploring and enhancing spatial thinking. *Current Directions in Psychological Science*, *22*(5), 367–373. https://doi.org/10.1177/0963721413484756

Voyer, D., Voyer, S., & Bryden, M. (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

Wei, W., Chen, C., & Zhou, X. (2016). Spatial ability explains the male advantage in approximate arithmetic. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00306