

Louisiana Tech University

## Louisiana Tech Digital Commons

---

Doctoral Dissertations

Graduate School

---

Summer 8-2021

### Exploring the Effect of Occlusion on a Computerized Mental-Rotation Test: Implications for Automatic Item Generation

Swadeep Patel

*Louisiana Tech University*

Follow this and additional works at: <https://digitalcommons.latech.edu/dissertations>

---

#### Recommended Citation

Patel, Swadeep, "" (2021). *Dissertation*. 941.

<https://digitalcommons.latech.edu/dissertations/941>

This Dissertation is brought to you for free and open access by the Graduate School at Louisiana Tech Digital Commons. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of Louisiana Tech Digital Commons. For more information, please contact [digitalcommons@latech.edu](mailto:digitalcommons@latech.edu).

**EXPLORING THE EFFECT OF OCCLUSION ON A  
COMPUTERIZED MENTAL-ROTATION TEST:  
IMPLICATIONS FOR AUTOMATIC  
ITEM GENERATION**

by

Swadeep Patel, B.S., M.A.

A Dissertation Presented in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

COLLEGE OF EDUCATION  
LOUISIANA TECH UNIVERSITY

August 2021

LOUISIANA TECH UNIVERSITY

GRADUATE SCHOOL

May 21, 2021

Date of dissertation defense

We hereby recommend that the dissertation prepared by

**Swadeep Patel**

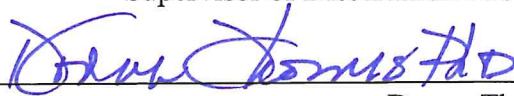
entitled Exploring the Effect of Occlusion on a Computerized Mental-Rotation Test:  
Implications for Automatic Item Generation

be accepted in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy in Industrial/Organizational Psychology**



Tilman Sheets  
Supervisor of Dissertation Research



Donna Thomas  
Head of Psychology and Behavioral Science

**Doctoral Committee Members:**

Steven Toaddy  
Jane Jacob

**Approved:**



Don Schillinger  
Dean of Education

**Approved:**



Ramu Ramachandran  
Dean of the Graduate School

## **ABSTRACT**

Radicals are characteristics that impact psychometric properties, such as item difficulty in tests. Within the mental-rotation test literature, occlusion and structure have been shown to contribute to item difficulty but require further study to be used within the context of automatic item generation (AIG). This study investigated whether high or low occlusion functions as a meaningful radical for Shepard-Metzler Mental Rotation Test (SMMRT) items. Thirty-two items containing computer-generated images of 3D block figures were administered to a sample of 180 participants on MTurk. After cleaning and removing careless responders from the sample, the data for 70 participants were analyzed using a 2 X 3 factorial ANOVA. Support for the hypotheses was not found; however, interesting responding patterns are observed due to high and low levels of stack occlusion. This pattern is further investigated in the discussion. The possible reasons for the response pattern are discussed, along with general recommendations and study limitations. Directions for future research are also provided.

## **APPROVAL FOR SCHOLARLY DISSEMINATION**

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It was understood that “proper request” consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author \_\_\_\_\_

Date \_\_\_\_\_

## TABLE OF CONTENTS

ABSTRACT.....	iii
APPROVAL FOR SCHOLARLY DISSEMINATION .....	v
LIST OF TABLES .....	vii
LIST OF FIGURES .....	viii
ACKNOWLEDGEMENTS.....	ix
CHAPTER 1 INTRODUCTION .....	1
Review of Literature .....	7
Computer-Based Testing and Automatic Item Generation.....	7
Cognitive Ability and MRTs .....	8
History.....	9
Shepard and Metzler .....	9
Vandenberg and Kuse.....	11
Item Difficulty .....	13
Problem Formulation .....	16
CHAPTER 2 METHOD .....	18
Materials and Procedure .....	18
Development of Testing Stimuli.....	19
Occlusion Measurement.....	21
Rotation Control.....	24

Analytic Plan.....	25
Main Study.....	25
Participants.....	26
Main Study.....	26
CHAPTER 3 RESULTS.....	27
Parametric Factorial ANOVA.....	30
CHAPTER 4 DISCUSSION.....	32
Power and Sample.....	32
Interaction and Opposites .....	37
Incentivization and Motivation.....	41
Angular Rotation when Considering AIG .....	45
Angular Rotation.....	45
Conclusion .....	50
REFERENCES .....	52
APPENDIX A DEMOGRAPHIC ITEMS.....	61
APPENDIX B HUMAN USE APPROVAL LETTER .....	63
APPENDIX C CUBE STACKS, OCCLUSION INDICATORS, AND SCORES .....	65
APPENDIX D MATRICES.....	68
APPENDIX E QQ PLOT FOR HIGH OCCLUSION.....	76

## LIST OF TABLES

Table 1	<i>Each Structure Type Presented Originally, Then with High and Low Occlusion</i> .....	21
Table 2	<i>Occlusion Scoring Example</i> .....	22
Table 3	<i>Table of Means and Standard Deviations for Occlusion and Structure</i> .....	29
Table 4	<i>Table of Means and Standard Deviations for Occlusion and Structure for the Data Including the 16 Artificially Generated Cases</i> .....	33
Table 5	<i>Scores Created for the Additional 16 Artificially Generated Cases</i> .....	34
Table 6	<i>Table of Means and Standard Deviations for Occlusion and Structure for the Data Including the 16 Artificially Generated Cases</i> .....	35
Table 7	<i>Table of Means and Standard Deviations for Occlusion and Structured With Inverted Mirrored and Different Scores</i> .....	38



## LIST OF FIGURES

<b>Figure 1</b>	Sample of Shepard-Metzler MRT pairs.....	5
<b>Figure 2</b>	Coordinate matrix and corresponding figure.....	6
<b>Figure 3</b>	Sample of Vandenberg and Kuse (1978) mental-rotations test. Line 1 answers: A and D. Line 2 answers B and C .....	11
<b>Figure 4</b>	Item structure descriptions and examples of presentation with minimal occlusion.....	16
<b>Figure 5</b>	Sample image of a stack (with and without block outlines).....	20
<b>Figure 6</b>	Box and whisker plot of occlusion scores for low and high occlusion items .....	24
<b>Figure 7</b>	Box and whisker plot of rotation scores for low and high occlusion groups .....	25
<b>Figure 8</b>	Mean scores plotted for structure by high and low occlusion. ....	31
<b>Figure 9</b>	Mean scores for structure by high and low occlusion where mirrored and different-type items had reversed outcomes. ....	39
<b>Figure 10</b>	Comparison of occluded figures in this study to the two established MRTs .....	44
<b>Figure 11</b>	GIF/image of a stack rotation. ....	46
<b>Figure 12</b>	Triaxial rotations for Euler angle calculation .....	47
<b>Figure 13</b>	Box and whisker plot comparisons between the first (top plot) and second (bottom plot) calculation methods for angular rotation scores. ....	49

## **ACKNOWLEDGEMENTS**

I would like to thank my dissertation committee chair, Dr. Tilman Sheets, for being such an incredible mentor and providing unwavering support during this project. I would also like to thank the other members of my dissertation committee, Dr. Steven Toaddy and Dr. Jane Jacob, for their service and support. My committee members always kept me pointed in the right direction.

# CHAPTER 1

## INTRODUCTION

During the past decade, online testing has dramatically increased in both organizational and academic contexts as test-takers can be located anywhere with a reliable internet connection (Cukusic et al., 2014; Reeves, 2000). However, un-proctored test administrations via the internet have raised many test-security concerns (e.g., using unauthorized materials, assistance from others, researching test questions online, recording and sharing test items) among test developers since its introduction (Foster, 2010). The concerns around test security include a long list of problems that revolve around the test-takers ability to cheat or aid in cheating, impacting how they score on the test (Karim et al., 2014).

Several security measures test administrators may consider ensuring the safety of their testing content (Tippins, 2015). These changes can be applied under two contexts: the test-administration process (proctored remote or in-person) or the test development/creation process. One example of the first approach locks the test taker's computer screen in an online remote-proctored context and requires them to record the room or testing setting before starting the test. During their session, they can be randomly asked to re-record the setting at any time, which lowers the odds of cheating. The second approach includes either having parallel forms of a test or having a method that reliably generates items automatically. This method helps with test security because it makes it

more difficult for test-takers to share items as test questions can be different during and between administrations (Cook & Eignor, 1991). Reliably generating items on the fly and parallel forms reduces the ability to cheat successfully because memorizing one instance of the test leading to cheating on the next instance is unlikely. After all, the items will not be the same.

One method by which items can be reliably generated is automatic item generation (AIG; Gierl & Haladyna, 2012). Using computerized AIG, a very large number of items can be generated quickly or even on-demand. The proper use of AIG methodology requires identifying the fundamental components of items used to assess a particular construct. To generate many items within the same construct space, the test developer must first identify and differentiate between the components of an item that are purely aesthetic and those components related to item difficulty. The aesthetic components are called incidentals and have no significant impact on item difficulty (Irvine, 2002). For example, in the operation of  $2+3$ , we can use  $3+2$ . This change does not necessarily change the item's difficulty but does allow for another similar item with which to assess a portion of mathematical ability. The components of an item related to the difficulty are called radicals (Irvine, 2002). Continuing our math example, adding a second rule of mathematics increases the difficulty (e.g., exponents; e.g.,  $2+2^3$ ). Through these examples, it is easy to identify in this context what we would think is a radical vs. an incidental; however, there are many other types of test items where the identification and definition of these components are much more difficult, such as in a mental-rotation test (MRT). Identifying and successfully manipulating these radicals are requirements for building a reliable and valid computer-adaptive AIG measure. For many modern testing

contexts, if future researchers want to build a computer-adaptive AIG measure, generating items that span the difficulty spectrum is necessary since the varied difficulty would be needed across items to measure the level of ability present in individuals specifically. Without varying levels of difficulty, computer-adaptive testing would not be possible.

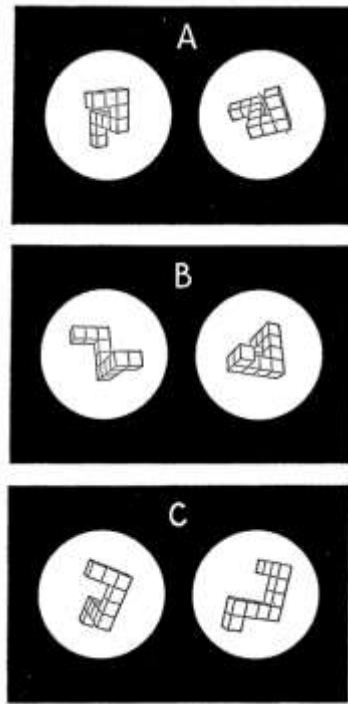
The MRT space is a ripe area for impactful AIG work. There are two major reasons why an MRT warrants more focus and should be employed in an AIG testing context. First, they are a non-verbal measure of cognitive ability, which has been described using the terms “culture-free” and “culture-fair” (Cattell, 1940; Cattell & Cattell, 1963). These tests, being mostly free from influences related to culture and language, can be used globally with limited need for translation other than simple instructions. Second, they are a well-established measure of cognitive ability. Studies have shown that cognitive ability is a strong predictor of success in academics and job contexts (Ones et al., 2006; Schmidt & Hunter, 2004).

Academic tests, some pre-employment assessments, and even measures of social intelligence are all indicators of general mental ability or cognitive ability (Spearman, 1927; van der Maas et al., 2014). Many theorists have conceptualized cognitive ability in a variety of ways (Gottfredson, 2002; 2004). Generally, researchers agree that cognitive ability is defined as an individuals’ general mental capacity to solve problems, plan, think in abstract ways, understand conceptually sophisticated and difficult ideas, and gain and learn new information quickly and efficiently (Gottfredson, 2004). It is easy to see why cognitive ability tests are used in employment and academic contexts. Employment-wise, they remain one of the most valid predictors of job performance; yet, some organizations

are rightfully hesitant to employ the use of these measures as they generally produce racial group differences (Campbell, 1996; Hulsheger et al., 2007; Schmidt & Hunter, 2004).

Cattell (1940) discussed that non-verbal cognitive ability measures eliminate language barriers and the chance for verbal loading, hence the term “culture-free.” Additionally, research supports the notion that they are less racially discriminatory than verbal measures of cognitive ability and tend to reduce adverse impact when used over global cognitive ability measures (Hausdorf et al., 2003; Ployhart & Holtz, 2008). Some examples of non-verbal cognitive-ability measures include Raven’s Progressive Matrices and MRTs. These tests all use various 2D and 3D shapes as a part of their items (Raven, 1938; Raven & Court, 1989; Shepard & Metzler, 1971; Vandenberg and Kuse, 1978).

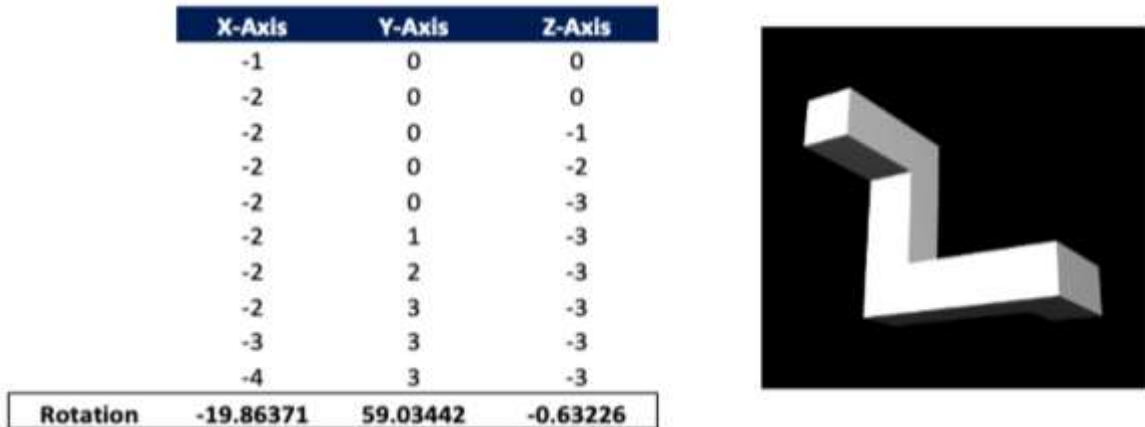
Specifically focusing on the SMMRT (1971), we can lay down the foundation for its use in an AIG methodology. When taking the SMMRT, the test-taker is presented with a pair of stimuli (Figure 1) and is instructed to look at the item on the left and determine if it matches the item on the right. The theory is that the test-taker engages in a mental-rotation exercise to match the two items.



**Figure 1:** Sample of Shepard-Metzler MRT pairs

Their test was simple yet elegant, and because of the simplicity and overall validity of this test, it is an excellent candidate to be used within an AIG context. Following very specific rules, an algorithm can populate a matrix consisting of the x, y, and z coordinates of each cube (10 coordinates in the case of this study) to replicate or generate items that resemble those created by Shepard and Metzler (1971). In Figure 2, each row specifies the starting coordinates of each cube in the stack. For example, the first cube in Figure 2 (bottom right corner) is positioned at -1 unit based on the cube's dimension, on the X-axis, and at 0 on both the Y and Z axes. The next cube added will move -1 unit from the first on the X-axis only. The third cube turns by following only the Z-axis 1 unit. The last row in the figure labeled *Rotation* measures how much the completed stack is rotated (in degrees) from the originally specified base point (i.e.,

0,0,0). Theoretically, the AIG would rely on an algorithm that uses item characteristics related to the difficulty to generate the ten coordinates to form a rotated cube stack and would allow for a test to be programmatically generated, providing convenience and test security along with the benefits of using a non-verbal measure of cognitive ability.



**Figure 2:** Coordinate matrix and corresponding figure

As originally proposed by Shepard and Metzler (1971), the items can be mirrored, rotated, or both mirrored and rotated. A later study introduces additional radicals: structural or “different-type” items (items that are not the same shape as the original, which fall into a similar category as mirrored) and the degree of occlusion of items (Caissie et al., 2009; Voyer & Hou, 2006).

Different-type items are those where the target stimulus image presented differs from the original stimulus image. Mirrored items fall into the same category as different-type items; however, with mirrored items, the target stimulus is a mirror of the first. Although the stacks appear to be identical, the images are mirrored and not the same. Each item structure contains some level of occlusion as it is impossible for a 3D figure not to block some portion of itself. In the current study, low-occlusion items are stacks



with an occlusion score of 0.59 or below. High occlusion items have been manipulated so that parts of the figure significantly block the other parts of the figure. High occlusion items will be those with an occlusion score of 0.62 and above.

This study will attempt to determine if the degree of occlusion (high or low) functions as a meaningful radical for SMMRT items. Using computer-generated images of 3D block figures (called “cube stacks” or “stacks”) similar to those used in the SMMRT, this study will examine the feasibility and effectiveness of utilizing figure occlusion as a radical in the creation of items in an MRT that future researchers can use for AIG systems.

## **Review of Literature**

### **Computer-Based Testing and Automatic Item Generation**

Smart devices have become an integral part of most workplaces. They are now more affordable, portable, and reliable than ever. The term smart device describes all devices such as phones, tablets, laptops, and desktops (Chernyshenko & Stark, 2015; Kozlowski & Bell, 2012). Due to the technology present in these devices, practitioners are no longer limited to the paper-and-pencil version of tests and test items. This technology enables test-takers to interact with unique items built from templates that share attributes related to item difficulty that adapt based on patterns of previous answers, which was never possible using traditional paper-and-pencil testing methods (Zenisky & Sireci, 2013).

An AIG methodology can produce many items using an algorithm and computer-based technology (Lai et al., 2016). Although there are several benefits of using an AIG approach, the main benefit is that AIG can quickly generate alternate forms using

specified parameters (Arendasy & Sommer, 2012). The ability to generate alternative forms also has implications regarding test security. With the classic testing examples (paper-pencil, un-proctored online, etc.), unauthorized copies of test items can be made public and easily accessible to anyone. However, with unique, automatically-generated tests, test takers can photograph and post items. The test administrator can rest assured that those items are unlikely to be used in any meaningful way in future administrations of the test due to very large item pools that can be created (Drasgow et al., 2009). Listed below are several practical improvements that AIG tests have over traditional test development:

- Reduces item-exposure concerns (Geerlings et al., 2011)
- Precise information on how the items were created, the items related to the construct being measured, and item psychometric properties are known
- Decreases investment of development time. Meaning, after initial development, AIG tests will produce an alternate form for each administration. In contrast, paper-pencil test items can be shared/compromised, and the test developer has to create a new test.

In the context of our SMMRT, we will be controlling the structure type of the item (different, same, or mirrored) and the degree to which the target stack is occluded. Both item structure and occlusion have been associated with item difficulty (Caissie et al., 2009).

### **Cognitive Ability and MRTs**

Cognitive ability is an individual's capacity to plan, adapt to new situations, and learn quickly (Neisser, 1967). Furnham (2008) adds reasoning, solving problems

(including thinking in an abstract manner), comprehending complex ideas, and quickly learning from experience to the list of domains included in the model. While there are theorists with far more complex models, there is one widely supported model that brings some level of uniformity to the overall construct—the two-factor theory (Guilford, 1967; McGrew, 1997; Spearman, 1904). Positive manifold was an observation made by Spearman (1927) where tests of cognitive ability, including academic measures and measures of social intelligence, were positively correlated, suggesting that there is an underlying variable that is responsible for this correlation. In other words, general intelligence within Spearman's two-factor model is theoretically the aggregate of all domain scores within the cognitive ability. This study will focus on non-verbal abilities, specifically spatial reasoning, within the broad aspect of general intelligence.

Shepard and Metzler (1971) and Vandenberg and Kuse (1978) MRTs are among the two most popular and well-known. Historically, MRTs have been used in neuroscience and neuropsychology to diagnose damage to the right cerebral hemisphere of the brain and the occipital lobe (Oostra et al., 2012). While this type of test is often used in the neuropsychology community to assess brain injury, this review will focus on an MRT using a non-clinical population focusing on better understanding item characteristics and its feasibility as an AIG test.

## **History**

### **Shepard and Metzler**

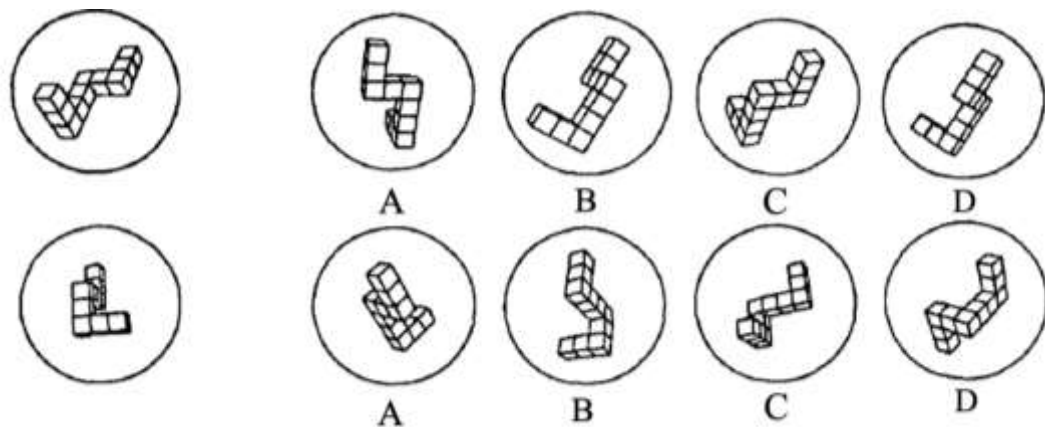
MRTs have long been linked to general intelligence based on the rate at which one can spatially process an image of an object (Jones & Anuza, 1982; Hertzog & Rypma, 1991; Johnson, 1990). Shepard and Metzler (1971) were among the first

researchers to study the mental-rotation-test space. Their initial experiment was primarily designed to test how long it took a participant to determine if the two objects presented were the same or not. Through this simple yet elegant experiment, they arrived upon some remarkable findings. This experiment showed the larger psychological community that thought processes were not only dependent on language but that analog representations also played a crucial role in these processes (Block, 1993).

In their ground-breaking 1971 study, Shepard and Metzler presented participants with images of a pair of asymmetrical cube stacks. They instructed them to decide if the two presented stacks were identical. The second stack was presented to the participant in a different orientation than the first stack. The second stack was either the same stack or a mirror image of the first. The authors hypothesized that the participants would create a mental representation of the first stack in their heads and rotate it to see if it matched the second stack in the image. Looking at Figure 1, the item in section A has identical stacks that are presented differently due to rotation of the z-axis, section B shows identical stacks that differ due to rotation of the y-axis, while section C shows mirrored stacks rotated similarly to section B. The researchers found that the amount of time it took individuals to determine if it was the same stack was related to rotation angle (measured  $0^{\circ}$ - $180^{\circ}$ ). Between participant differences in timing were related to individual differences such as test scores of cognitive ability and reaction time measures. The relationship between item reaction time and angular rotation was still linear—the greater the degree of change in angle of rotation (regardless of across which axes it was rotated), the longer the reaction time. This correlation was consistent whether participants practiced or not.

### Vandenberg and Kuse

In 1978, Vandenberg and Kuse developed their measure of spatial visualization that was similar to that presented in the 1971 Shepard and Metzler study. Their original test consisted of 20 items of printed cube stacks. Each item consisted of an original stimulus and presented four answer options (two distractors and two correct items; see Figure 3). Items were presented to test-takers in four sets of five. Half of the items contained two rotated mirrored distractors, while the other half contained two distractors with different rotated stacks recycled from other questions. The total amount of time given to complete the test was 10 minutes. The test was administered in two halves (if they wanted a test-retest correlation), and five minutes were allowed for each half.



**Figure 3:** Sample of Vandenberg and Kuse (1978) mental-rotations test. Line 1 answers: A and D. Line 2 answers B and C

The measure was administered to various age and educational-level groups (university, high school, and elementary school) across three years. Results suggest that the Vandenberg and Kuse MRT can be used in studying the development of spatial ability. These findings were interesting, especially when the general belief was that sex differences in spatial ability/mental rotations increase at the beginning of puberty

(Vandenberg, 1975). However, in the scope of cognitive-ability measurement, the 1978 study shows that the MRT is moderate to strongly correlated with other measures of spatial ability, with Pearson correlations ranging from 0.39 to 0.68. The 1978 Vandenberg and Kuse MRT also showed very weak correlations with measures of verbal ability, with Pearson correlations ranging from 0.03 to 0.07.

Logically, there is an alternate explanation of the weak correlations between MRT and measures of verbal ability. As previously mentioned, Cattell and Cattell (1963) established that non-verbal measures tend to be “culture fair.” Therefore, the authors concluded that the introduction of confounds like culture in measures of verbal ability could help explain why these two different measures of cognitive ability are not highly related. They propose that a larger portion of the variance in the scores of verbal measures might be better explained through the lens of culture rather than a measure of cognitive ability (Cattell & Cattell, 1963; Horn & Cattell, 1966).

Many theorists around this time did not accept the idea of mental rotation. Just and Carpenter (1971, 1975), Hochberg and Gellman (1977), Steiger and Yuille (1983), and Marks (1999) were concerned with how the SMMRT images are represented and manipulated within the mind. These researchers and others decided to hold a microscope to the linear function found by Shepard and Metzler. Carpenter and Just’s (1978) work with tracking eye movements suggested that the increase in time was not due to the mental rotation of the stimulus but the need to make more movements between the two stimuli to compare the item features. Their main criticism of MRT's idea is that holding the items side-by-side does not methodologically lend enough evidence that individuals form mental representations of these items and manipulate them in their minds; rather,

they compare item features in a piecemeal fashion to determine if the figures are same. The more comparisons they have to make, the more time it will take, thus offering an alternative explanation for the linear relationship observed by Shepard and Metzler (1971).

Within the MRT literature, side-by-side feature comparison vs. mental rotation remains one of the most controversial topics. The underlying mechanics of mental imagery have been the focus of study by several neuroscience researchers. Georgopoulos et al. (1989) studied MRT tasks in a rhesus monkey using neural implants and supported the mental-rotation hypotheses. Other researchers used functional magnetic resonance imaging (fMRI) to study MRT (Koshino et al., 2005; O'Boyle et al., 2005; Richter et al., 2000). Richter et al. (2005) used the SMMRT and found involvement of brain areas that are most likely to participate in mental rotation (superior parietal lobule and lateral premotor area). Although neuroscientific evidence is supportive of the mental-rotation task, this area is still debated (Carpenter & Just, 1978).

### **Item Difficulty**

Based on the review of the literature of MRT, the researcher believes that three main questions remain to be answered because either there is not enough research or research is highly controversial, and results are often debated.

1. Does an MRT require mental rotations? (Contested results)
2. Are there gender differences in MRT scores? (Controversial with a mixed bag of results)
3. What are the item features related to the difficulty of MRT items? (Needs further study)

The primary focus of this study is to shed more light on the third topic: the identification of item characteristics related to the difficulty of MRT items.

Identifying the item features related to difficulty is necessary to understand each feature's contributions better. It allows researchers to manipulate those features in the creation of new items. Previous studies have identified several possible contributors to the difficulty of items, including the following features, which can be entailed individually and/or in combination in MRT items: occlusion, specific item structure (mirrored, different, or same), and configuration of peripherals. As mentioned earlier, occlusion is the degree to which one part of a figure blocks the view of another part of the same figure, thus relying on the observer to make an inference of what is blocked. Mirrored items are those wherein the original stimulus is reflected across a single plane or axis; in this sense, the two stacks are technically different in this single and extremely constrained manner. Different-type items are those where the target stimuli are a different stack than the original stimuli in any other way than the mirroring described above. Unlike different types of items, mirroring gives the test-taker a convincing illusion of the same figures (Caissie et al., 2009, Shepard & Metzler, 1971; Voyer & Hou, 2006).





Additionally, Caissie et al. (2009) found that occluded items were more difficult than items not occluded. Items that were mirrored were more difficult than items that were structured differently. They also listed other factors contributing to the difficulty that are not a physical part of the items *per se*, including the amount of time allotted per item. The introduction of a time limit adds pressure that forces the participant to engage in mental rotation rather than solve the puzzle by other methods (e.g., counting blocks, looking back and forth, etc.). The suggested time limit for Vandenburg and Kuse is 10



minutes for the 24-item measure. The Shepard and Metzler Method does not have a suggested limit as it only measured time to respond and did not force a response within a certain amount of time.

Item structure and degree of occlusion are the variables of interest for this study. However, studying structure under the Vandenberg and Kuse MRT paradigm is more complicated, as participants are presented with five cube stacks per item (one original stimulus and four target stimuli). Additionally, mirroring was the only different-type distractor used. Analyzing those data would require many more participants in order to obtain acceptable statistical power. However, under the SMMRT paradigm, less power would be required, and isolating the role of occlusion would be less complicated.

This study will attempt to look at the impact that low and high levels of occlusion have on the difficulty of MRT items. Mirrored items, different-type items, and same-type items will be randomly presented to participants, with each type of item being presented in low occlusion or high occlusion configurations. The total rotation of each stimulus (the sum of the absolute value of the rotation on each axis) used in the study was calculated. To assure that occlusion and rotation are independent of one another. The total rotations were then grouped based on high ( $>0.62$ ) and low ( $< 0.59$ ) occlusion items and then compared using a t-test to determine if higher occlusion items had significantly higher rotation scores. If so, this would mean that angular rotation is a major confound. See Figure 4 for examples of the items used in this study.

Original Stimulus	Manipulated Item	Description
		<p><b>Mirrored:</b> Original stimulus is reflected across a plane to create a mirror figure.</p>
		<p><b>Different-Type:</b> Item is a visibly different structure than the original stimulus.</p>
		<p><b>Same-Type:</b> Item and stimulus are identical. It will be shown from a different angle.</p>

**Figure 4:** Item structure descriptions and examples of presentation with minimal occlusion

### Problem Formulation

As the research currently stands, there is no standard approach to manipulating occlusion within MRT-type items. This study will focus on the item characteristic of occlusion. It is a reasonable option among several item characteristics (item type, time, etc.) and associated with item difficulty (Caissie et al., 2009; Voyer & Hou, 2006).

Caissie et al. (2009) posit that occlusion contributes to difficulty because it increases the cognitive load regardless of the three structure types. For practical reasons, this study utilized items having low or high levels of occlusion within each structure type to determine the impact occlusion has on difficulty. The items were categorized as low and high when selected and then checked using an occlusion scoring method. The occlusion scores for the items formed a bimodal distribution and have varying degrees within the high and low groups. Future researchers can use the occlusion scoring approach presented in this study to test the veracity of a continuous version of occlusion and its relationship to difficulty and deploy an AIG version of SMMRT.

*Hypothesis 1:* High levels of occlusion will be associated with higher levels of difficulty than lower levels of occlusion regardless of the structure of the item (mirrored, structural, same).

In conformity with the findings of Voyer and Hou (2006), Caissie and colleagues found that the mirrored types of items were more difficult than items that were a different structure. Mirrored items, being deceptively similar to the original stimulus, is the reason for their high difficulty.

*Hypothesis 2:* Mirrored-type items will be associated with higher levels of difficulty than other structure types (mirrored, different, same).

The purpose of this study is fourfold. The first is to replicate previous findings as they relate to item features contributing to the difficulty. The second is to study occlusion in two varied states (low and high). The third is to introduce a method that allows researchers to measure occlusion levels which can then be used to study occlusion as a continuous variable. Lastly, this study will address implications for applying this method to an AIG context.

## **CHAPTER 2**

### **METHOD**

#### **Materials and Procedure**

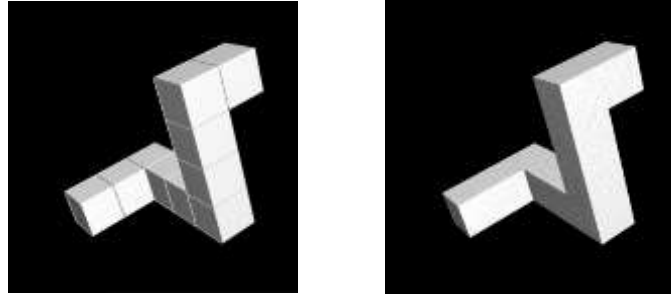
Participants were administered a computerized version of SMMRT where each item was randomly presented without replacement via the Qualtrics survey platform (see Appendix A for full questions list). The mental-rotation measure consists of 32 total items. Participants were presented with two stimuli of cube stacks and were asked if each of the two stimuli are different views of the same underlying stack. The items in the measure utilized six highly occluded stacks and 16 minimally occluded stacks. Each group of highly and minimally occluded items contains four items: mirrored and different stacks and eight “same-type” stacks. Same type items were doubled to equalize the number of yes and no responses; however, only the first half will be scored. Thus, the measure has 24 scored questions out of 32.

Participants were encouraged to answer the questions quickly and accurately (see Appendix B for instructions) within the allotted time. Monahan et al. (2008), using a computerized version of an MRT, allowed participants three minutes per block of 12 items, averaging 15 seconds per item, and allowed participants to spend more time on more difficult items and less time on less difficult items. The 15-second limit used in this study was intended to discourage participants from attempting to engage in nonmental-rotation activities (e.g., back-and-forth figure comparison) that Carpenter and Just (1978)

and other researchers sometimes observed. Once participants have completed all 24 SMMRT items, they were forwarded to a final page that thanked them for their participation (see Appendix B).

### **Development of Testing Stimuli**

The cube stacks were created using an R script that extensively uses the RGL package (Sheets, 2020). The RGL package allows users to design and generate 3D figures based on a few parameters. The R code used in this study allows the user to rotate each figure with a high degree of precision. The code builds cube stacks that can be manipulated to change the radicals as needed (e.g., occlusion, rotation), then be saved as an image. In addition, the build and rotation data for each stack can be saved to a spreadsheet and used to generate identical stack images by others. For this study, four stacks were developed using the following rules to add consistency. First, all stacks contained a total of 10 cubes. Second, all stacks contained a total of four legs or three right-angle bends in the stack. To account for the findings of Carpenter and Just (1978) and Caissie et al. (2009), where there is evidence of test-takers comparing figure features rather than engaging in mental-rotation, I removed the block outlines for each stack, so it becomes one consistent element rather than stacked cubes, thus making it more difficult for participants to count (see Figure 5).



**Figure 5:** Sample image of a stack (with and without block outlines)










The first type of items created was the *same-item* type, where the original cube stack was rotated to include high occlusion, then rotated to include minimal occlusion. To create the additional eight items that would equalize the correct/incorrect responses, the original stimulus and target stimulus for the eight items that were first generated were swapped so that the original stimulus is presented as the second item. These items were presented after the original eight were presented to minimize practice effects, and responses to these items were excluded from the analysis to maintain balanced groups. The total degrees of rotation between the low and high occlusion were calculated after the items were developed and were found to have similar degrees of rotation between both groups. Next, a mirror function was used to produce a mirrored image of the original cube stack. Then the stack was rotated to include high occlusion, then rotated to include minimal occlusion. Finally, to create the different-type item, cube stacks that were structurally different from the original cube stack were created and rotated to include high occlusion, then rotated to include minimal occlusion.

Contrary to different type items, mirrored items require that each feature is held constant except for one. The easiest are noticeably different within the difficulty spectrum for different-type items (bends in the same dimensions, different number of cubes on the ending legs, etc.), and most difficult are the mirrored items. Many variations

create the different types, while the mirrored is created with just one type of manipulation. This method was repeated to all six of the original cube stacks to form an MRT of 24 items. Table 1 shows all item types, and Appendix C shows all the stacks created using RGL (see Appendix D for the stack matrices).

**Table 1**

*Each Structure Type Presented Originally, Then with High and Low Occlusion*

	<u>Original Stack</u>	<u>High Occlusion</u>	<u>Low Occlusion</u>
Mirrored Type			
Same-Type			
Different-Type			

### Occlusion Measurement




A search via EBSCO of scientific databases provided no reference related to a methodology to measure item occlusion. One of the main contributions of this study is to provide future researchers using SMMRT as a method of occlusion measurement. The background is black, while the cube stacks are white and gray, using the stacks created for this study as an example. The contrast allows a measure of the degree to which the image in the foreground (white and gray; any HTML color code > 0) is separated from its background (black; HTML color code = 0). The background is always black, while parts

of the figure with clear visibility are colored using a spectrum of pixels from white to dark gray.

This study measured occlusion by comparing the number of black pixels to white and gray pixels. The image size was kept consistent (400x400 pixels = 160,000 total pixels). Subtracting the white and gray pixels from black and dividing them by the total pixel count provides a score for occlusion, where high scores indicate high levels of occlusion. To future researchers, this score will allow for a more accurate measurement of occlusion. See Table 2 to see how the score is an indicator of occlusion. Both the *Original Stack* and *Low Occlusion* columns give figures where there is very low occlusion. There is a noticeable difference between the black and white/gray pixel counts compared to the high-occlusion figure. When looking at the occlusion score in Table 2 and Appendix C, we can see that items with low occlusion range from 0.46 to 0.59, with high levels of occlusion range from 0.62 to 0.78.

**Table 2**

*Occlusion Scoring Example*

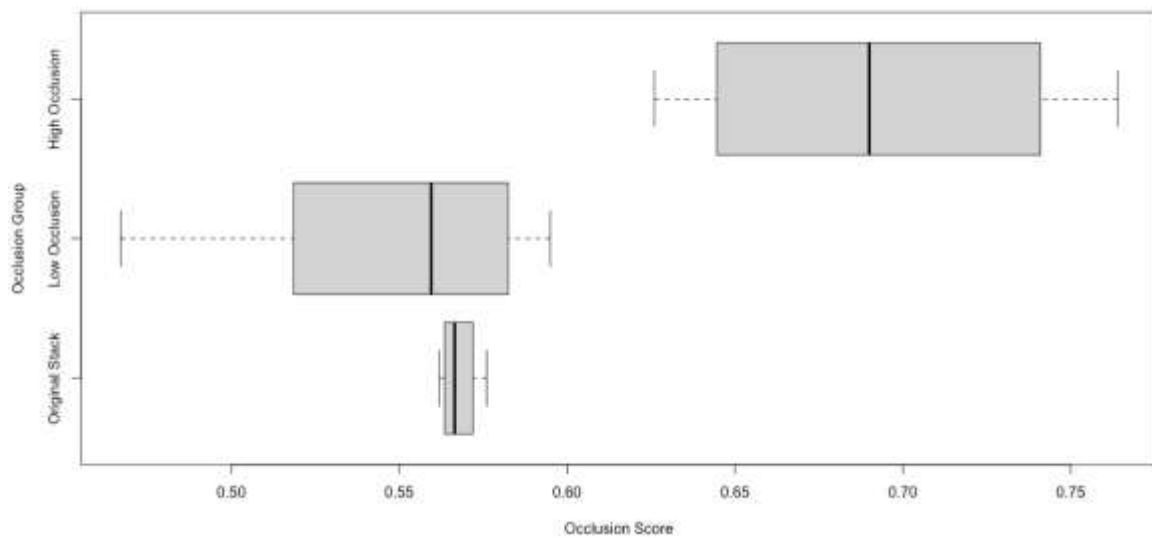
<u>Original Stack</u>	<u>High Occlusion</u>	<u>Low Occlusion</u>
		
● Black: 125025	● Black: 141023	● Black: 129445
● White/Gray: 34975	● White/Gray: 18977	● White/Gray: 30555
● Occlusion Score: .562	● Occlusion Score: .762	● Occlusion Score: .580



To further study the occlusion score, the means and SDs were calculated for the original, low, and high stacks. The average occlusion score for the original four stacks is 0.568, with an SD of 0.006. The mean occlusion value of the 12 low-occlusion stacks is 0.549 with an SD of 0.04, while the high-occlusion average for the 12 high-occlusion stacks is 0.689 with an SD of 0.05. Since the original stacks are considered low occlusion items, we see similar occlusion scores.

The box-and-whisker plot in Figure 6 shows a clear demarcation in scores between low and high occlusion items. For this study, every stack equal to or less than 0.59 was a stack categorized as having lower levels of occlusion. In comparison, any stack equal to or greater than 0.62 was a stack categorized as having higher levels of occlusion. A *t*-test was conducted to determine further if there were significant differences in the scores in the high and low occlusion groups. The *t*-test resulted in  $t(26) = -8.44$ ,  $p < .001$ , with Cohen's  $d = 0.42$  suggesting a significant difference between the high and low occlusion groups with a medium effect.

Future researchers that would adopt this or a similar method of calculating occlusion score would also need to determine how to calculate each image's absolute center and control for image zoom. In this study, even without maintaining that degree of rigidity, the occlusion scores that were calculated lined up properly with the predetermined groupings of low and high occlusion. There was a clear distinction in measurement between low and high occlusion that can be studied linearly in the future.



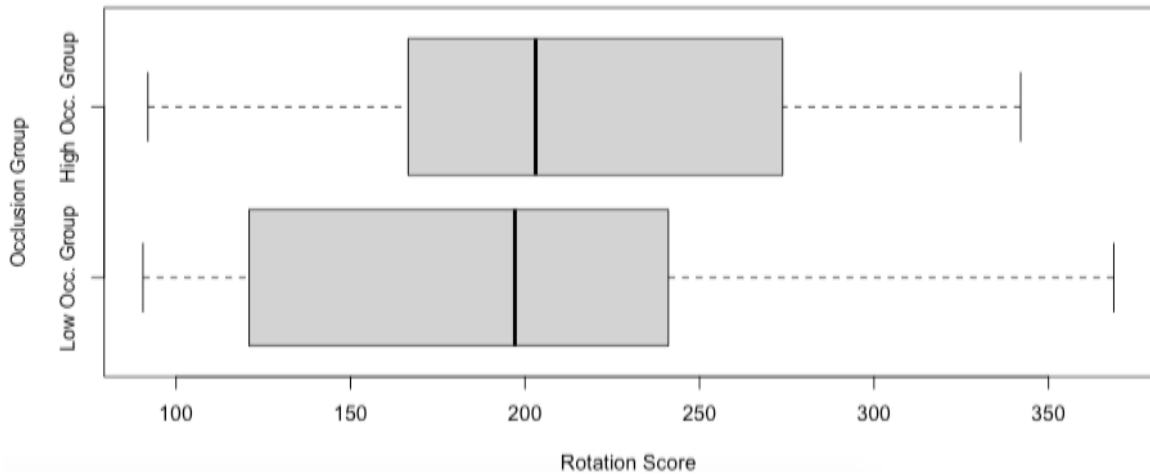
**Figure 6:** Box and whisker plot of occlusion scores for low and high occlusion items

### Rotation Control

The total degrees of rotation for each item was calculated to ensure that the degrees of rotation and occlusion are independent. This calculation took the absolute value of each degree of rotation by axis (since a negative rotation is still a rotation) and was summed to arrive at a total rotation score. Meaning the rotation in degrees for each of the X, Y, and Z axes were summed. A t-test was used to compare the total rotation scores between low and high occlusion groups to see if rotation between these groups significantly differed. A significant difference in rotation scores would indicate that rotation and occlusion are not independent of each other. Ideally, the rotation scores in each group should be equal.

The low and high occlusion groups each had 12 scores. For the low occlusion items, the average total rotation score is 198 with an SD = 83.7, while the average for the high occlusion items is 212 with an SD = 73.6. Levene's test ( $p = 0.66 > p = 0.05$ ) was not significant so equal variances are assumed. The  $t$ -test resulted in  $t(22) = -0.44$ ,

$p^o=0.66$ , which fails to indicate significant differences in rotation between low and high occluded items. See Figure 7 for a box and whisker plot for low and high occlusion groups.



**Figure 7:** Box and whisker plot of rotation scores for low and high occlusion groups

## Analytic Plan

### Main Study

For this study, occlusion and structure type are the item features of interest or independent variables (IV). The outcome measure or the dependent variable (DV) will be difficult, which the number of incorrect responses will measure.

Our first IV, occlusion, has two (high and low) levels, while the type has three levels (mirrored, same, different). Since there are two IVs and one DV, a factorial or two-way ANOVA analysis (ANOVA) is the recommended method to test the hypotheses. This ANOVA should yield two significant main effects. The first main effect will be occlusion on difficulty (H1), while the second will be that of structure type on difficulty (a necessary precondition for H2).

A significant main effect of occlusion on difficulty and a lower mean score (meaning a low number of total items correct) for items with high occlusion will need to be observed to support H1. To find support for H2, a significant main effect of structure type on difficulty and a lower mean score for mirrored items would have to be observed. Also, Tukey's HSD post-hoc test needs to be conducted, and mirrored items should form the lowest homogeneous subset. Both same- and different-type items would either be similar or in separate, higher homogenous subsets in either order.

## **Participants**

### **Main Study**

The number of participants required for this study was determined using the g\*Power 3.1 power analysis software. It was determined that 86 participants will be needed to detect a medium effect size at  $\alpha = 0.05$  with power = 0.80. An Amazon MTurk sample was used to obtain participants for this study. Due to a miscalculation in the original power analysis, concerns regarding potential data quality issues, and other factors associated with un-proctored online testing (e.g., incomplete responses, clicking through the measure without paying attention, technical issues on the test-taker end, etc.), the sample size was increased to 180. Upon cleaning the data and removing participants who carelessly responded, the study's number was 70. The method for identifying and removing the problematic data will be covered in the next section. Lastly, the reason for not collecting additional data to reach the ideal sample size will be discussed in Chapter 4.

## **CHAPTER 3**

### **RESULTS**

The data were cleaned, formatted, and explored for missing responses and guessing. Out of the 180 cases collected from MTurk, 16 cases showed signs of guessing where participants either finished the measure in less than one minute (estimated time for completion was between 6-10 minutes) or answered all questions with the same response. A total of 164 cases were retained. Next, exploratory analysis, including checking assumptions for a two-way ANOVA, was conducted according to Field (2009). The homogeneity of variance assumption was not met, and normality plots appeared to be extremely positively skewed. A high concentration of low scores was interpreted as a possible indicator that other cases where guessing and careless responding could occur could be a reason for the failed assumptions and further investigate and clean the misleading data. Finally, the participant's means and SDs for response time were calculated to get insight into the dispersion. The reason being, ideally, candidates should have similar response times across the measure, possibly varying 2-3 seconds, and not double the average response time. A high SD for response time served as an indicator that the participant may have rushed through by providing careless responses to some questions and then timed out by not paying attention to others.

In the process of building this measure and testing its functionality, it was determined that it is highly unlikely for an individual to respond to under two

seconds. The time it takes to look at the stacks, process them, and move the mouse/pointer from the answer choice to the “next” button would have lapsed a minimum of two seconds. Using this, as a rule, 61 participants had an average response time of two seconds or below, while 54 responded to more than one item in one second or less. These participants were excluded from the sample leaving a total of 103. Next, an additional eight participants with average response times between 2-3 seconds were investigated. These participants had a wide range of response times where their SDs were more than double their average response time. Upon further investigation, it was determined that they responded to some questions in under one second and timed out in 15 seconds for the others. Since there was no consistency in their response pattern, they were excluded from the sample. Lastly, looking at the top end of the spectrum (responses at timing out at 15 seconds), there were 25 participants that either left their computer while taking their measure or intentionally let it time out to increase their average response time on the HIT. Unusually quick and unusually lengthy times for within-participant data have been cautioned by a few authors as evidence for inattentiveness (Hauser et al., 2018; Kittur et al., 2008). As these are coded as incorrect responses, and a large portion of their responses timed out, this severely impaired their score and the distribution of the overall dataset, which raised the question of if they were actually trying and just timed out at an unusual rate, or if they did not pay attention and carelessly took the measure. To follow a more conservative approach that would avoid including people that did not try, these 25 participants were excluded from the sample leaving a remainder of 70 participants in the sample. After the data were cleaned, the means and SD for the 70 participants were calculated. The average time spent on the measure for

these cases was 6.62 minutes with an SD of 4.1 minutes, congruent with the expected time for completion. The average number of correct responses for the overall measure was 13 out of 24 (55%), with an  $SD = 3.9$  (16%). See Table 3 for means and standard deviations by item type.

Table 3

*Table of Means and Standard Deviations for Occlusion and Structure*

<u>Occlusion Level</u>	<u>Structure Type</u>	<u>Mean Score</u>	<u>Standard Deviation</u>
High Occlusion	Mirrored	2.57	1.14
	Different	2.49	1.18
	Same	1.59	1.1
Low Occlusion	Mirrored	1.51	1.32
	Different	1.76	1.26
	Same	2.93	1.35

Next, the assumptions for ANOVA were tested. The Kolmogorov-Smirnov and Shapiro-Wilks normality tests were again significant, suggesting that the data were not normally distributed. However, the QQ plots looked acceptable as all plotted points were extremely close to or overlapped with the 45-degree reference line (See Appendix E). Levene's test based on the mean was significant at  $p < 0.001$ ; however, the test was nonsignificant based on the median ( $p = 0.135$ ). Since the sample sizes for each level of the IV are equivalent, the analysis can be continued according to Stevens (1996) due to ANOVA being robust against these violations of normality and homogeneity of variance; hence, this study will present the results parametric factorial. Generally, nonparametric robust factorial ANOVA would be the suggested method for analysis when data do not follow a specific distribution. Two methods for nonparametric analyses were explored

before coming back to the parametric analysis: Procedures for an Aligned Rank Transform (ART) ANOVA and a robust nonparametric ANOVA (Wilcox, 2005; Wobbrock et al., 2011). Both alternatives presented distorted findings due to mean trimming and transformations. For example, the robust method excluded scores for cases when the case had a 0 or 100. While this would have worked for other data sets with a more continuous DV and significant outliers, this method was not applicable for these data.

### **Parametric Factorial ANOVA**

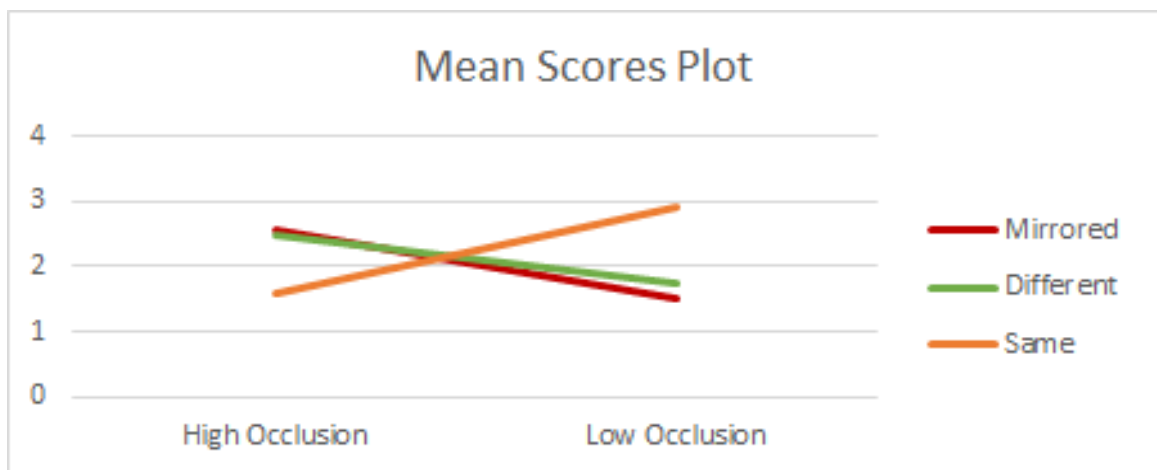
The results for the ANOVA showed nonsignificant main effects for both occlusion and structure type. Both effects are necessary to support both of the hypotheses for this study. As it stands, neither occlusion nor structure type has an impact on difficulty:  $f(1,414) = 1.7, p = 0.195$ , and  $f(2,414) = 1.2, p = 0.299$ , respectively. The interaction term, however, was significant  $f(2,414) = 43.61, p < 0.001$ , partial eta squared = 0.174, indicating a large effect.

The hypotheses for this study were rejected since an interaction term is present. According to Field et al. (2012), the recommended path would be to ignore main effects in the presence of a higher-order effect. The first hypothesis (i.e., H1) focused on occlusion's impact on difficulty regardless of structure. At the same time, H2 expects mirrored items to be the most difficult, making them both dependent on a significant main effect. There are several cases where the interaction term would be worth exploring. For example, a possible finding that would lend support to H1 would include occlusion always being harder for one type of item and only somewhat harder for other item types. For H2, in this hypothetical scenario, mirrored items would always be more difficult than



the same or different type items. However, high-occlusion mirrored items would be more difficult than low-occlusion mirrored items. However, the findings do not support this.

When looking at the raw mean scores (Table 3), there is evidence that the exploration of the interaction term is not meaningful. Mirrored items do not always have the lowest mean score. The mean score of all high occlusion items is higher than that of low occlusion items, indicating that lower occlusion was more difficult (see Figure 8).



**Figure 8:** Mean scores plotted for structure by high and low occlusion

## **CHAPTER 4**

### **DISCUSSION**

#### **Power and Sample**

There are two reasons additional data were not collected to reach the recommended sample size; the first data collection is based on statistics. The second data collection is practice-based. The *p*-values and effect sizes reported for the interactions found were statistically significant with a large effect. It is highly unlikely that collecting an additional 16 responses would change the trend. Secondly, the practical reason, due to the amount of attrition associated with obtaining 70 participants, data from an additional 45 participants would need to be collected in hopes of ending up with 16 participants. Ultimately, the second reason ties back into the first. It would not have made statistical or practical sense to do so. Two tests were conducted to determine if an additional 16 participants would have been necessary. The first was to bootstrap data for 16 participants in SPSS and then performed the analysis. The second was to purposefully create 16 participants that would provide the results that support the hypotheses to see what the results would be like if 16 ideal participants were added.

A stratified bootstrapping was conducted using SPSS to generate data for the 16 participants. The mean for the bootstrapped sample was 2.38, with an SD of 1.36. A *t*-test was conducted to determine if the sample means were significantly different. Levene's

test was not significant, indicating that the equal variance assumption is met. The  $t$ -test was not significant  $t(514) = -1.6, p = 0.11$ , indicating that the means between the original sample and bootstrapped sample were not significantly different. The bootstrapped data were added to the original data for hypothesis testing. The QQ plots were identical to the ones referenced in the results section. Levene's test was again significant, indicating that equal variances cannot be assumed. Since ANOVA is robust against this assumption violation, the analysis was still conducted. Both the occlusion and structure main effects were not significant  $f(1,510) = 0.862, p = 0.353; f(2,510) = 0.299, p = 0.741$ . The interaction term was again significant  $f(2,510) = 53.11, p < .001$ , with partial eta squared = 0.172 indicating a significant interaction term. The means and SDs for the sample are in Table 4.

Table 4

*Table of Means and Standard Deviations for Occlusion and Structure for the Data Including the 16 Artificially Generated Cases*

<u>Occlusion Level</u>	<u>Structure Type</u>	<u>Mean Score</u>	<u>Standard Deviation</u>
	Mirrored	2.61	1.12
High Occlusion	Different	2.56	1.16
	Same	1.52	1.10
	Mirrored	1.65	1.36
Low Occlusion	Different	1.81	1.31
	Same	2.94	1.01

For the same reason as the original analysis, post-hoc would not be meaningful as the mean for high occlusion is higher than low occlusion, and mirrored items are not the lowest mean in both groups.

For the second method, a sample was created following specific rules that would support the hypotheses. For H1, the average score for high occlusion must be lower than the average score for low occlusion. The mean for all high occluded items was two points lower than the mean for all low occluded items. For H2, the average score for mirrored items has to be lower than the same and different in both high and low occlusion groups. The average score for mirrored items was two points lower than the average of same and different type items in both occlusion groups. Below is a table of the scores for each item type. These scores were repeated 16 times to create data for 16 participants.

Table 5

*Scores Created for the Additional 16 Artificially Generated Cases*

<u>Occlusion Level</u>	<u>Structure Type</u>	<u>Score</u>
	Mirrored	0
High Occlusion	Different	2
	Same	2
	Mirrored	2
Low Occlusion	Different	4
	Same	4

The same steps conducted to analyze the 70 cases discussed in the results section were followed to analyze these 86 cases. The QQ plots looked similar to those reported

above; however, Levene's test was statistically significant. Again, due to the robust nature of the ANOVA, the decision to move forward was made. The main effect of occlusion was significant  $f(1,510) = 5.66, p = 0.018$  with a partial eta squared of 0.011, indicating a small effect. The main effect of structure was significant  $f(2,510) = 9.93, p < .001$ , with a partial eta squared of 0.037 indicating a small effect. The interaction term was significant  $f(2,510) = 33.35, p < .001$ , with a partial eta squared of 0.116 indicating a medium effect. See Table 6 for the means and SDs for item and structure type.

Table 6

*Table of Means and Standard Deviations for Occlusion and Structure for the Data Including the 16 Artificially Generated Cases*

<u>Occlusion Level</u>	<u>Structure Type</u>	<u>Mean Score</u>	<u>Standard Deviation</u>
	Mirrored	2.09	1.44
High Occlusion	Different	2.40	1.08
	Same	1.66	0.98
	Mirrored	1.60	1.20
Low Occlusion	Different	2.17	1.43
	Same	3.12	1.00

Adding 16 ideal cases (roughly 18% more of the sample) did change the results towards supporting my hypotheses. Using extreme cases had the desired impact on the means scores for the groups and shifted them towards the study hypotheses. However, since a significant interaction term was found, the simple effect post-hoc analyses test was conducted. First, an occlusion constant was held and examined along with each level of occlusion. The univariate tests for high and low occlusion showed that both groups

were statistically significant, meaning that the scores for structure in each condition were statistically significant, which indicates that levels of occlusion did have an impact on item difficulty for some structure types; however, the direction of difficulty can be extrapolated from Table 6. Pairwise comparisons showed that the same-type item was significantly more difficult in high occlusion groups than the different-type item. Pairwise comparisons showed that mirrored-type items were significantly more difficult than the same type in the low occlusion group. Different-type items were significantly more difficult than the same type, suggesting that mirrored and different items were more difficult in low occlusion and not in high occlusion, and the same items were difficult in high occlusion but not in low occlusion. The conclusion for these findings is similar to the original and bootstrap sample.

For the next comparison, the structure was constant, and occlusion was investigated at each structure type. The univariate test showed that mirrored and same type items were statistically significant, but not different types, indicating that the high and low occluded items had statistically different means for the two structure types. In the mirrored condition, low occluded items were significantly more difficult than high occluded items. In the same-type condition, high occluded items were significantly more difficult than low occluded items. It suggests that mirrored items were only the most difficult type of item in the low occlusion space.

In contrast, the same type was the most difficult in the low occlusion space. These conclusions do not fully support both H1 and H2. Again, the conclusions from this sample are similar to the original and bootstrapped sample; therefore, the decision to collect real participant data to get to a sample of 86 was not pursued.

### **Interaction and Opposites**

Based on the data analyses, support was not found for either H1 or H2; however, an interesting pattern was discovered during cleaning and analysis. First, after cleaning the data and drawing a reasonable conclusion that the data and participants retained did make an honest effort to assess, an opposite responding pattern was observed for mirrored and different structured items based on occlusion. This responding pattern was opposite to the hypotheses. We expected low mean scores for high occlusion and high mean scores for low occlusion, but the data displayed the opposite. Based on this finding, it is a conjecture that the level of occlusion impacts participants' response strategy. The second finding is that there was a significant amount of guessing and careless responding among the participants, reducing the sample from 180 to 70. Possible motivations and reasons for guessing will be explained in the subsequent section.

Figure 8 shows that participants were more inclined to say that the stacks were not the same (resulting in correct responses for mirrored and different and incorrect for same). In contrast, for low occlusion items, participants were inclined to say the figures were the same. If so, it may be that occlusion levels impact participants' response strategy, so much so that there seems to be a consistent pattern when high occlusion items are presented. Suppose it was difficult for participants to discern the overall similarity of items due to high occlusion. In that case, they might have decided that answering “no” would be safer, and it might be the case if they think that the measure tries to “trick” them when presented with a highly occluded stack.

In addition, it appeared that when looking at low occlusion items, the participants showed a consistent pattern of answering “yes.” The correct/incorrect outcome for

mirrored- and different-type items were reversed to test this hypothesis so that the focus of the analysis would be whether they provided a yes or a no. The same type of items was not manipulated as the response format for those items was already correct, allowing us to research the response pattern further. This manipulation is only limited to this analysis. Table 7 shows the means and SDs after the scoring paradigm was changed for mirrored and different items. See Figure 9 for the plotted means of the three-item types by occlusion.

Table 7

*Table of Means and Standard Deviations for Occlusion and Structured with Inverted Mirrored and Different Scores*

<u>Occlusion Level</u>	<u>Structure Type</u>	<u>Mean Score</u>	<u>Standard Deviation</u>
High Occlusion	Mirrored	1.43	1.14
	Different	1.51	1.18
	Same	1.59	1.07
Low Occlusion	Mirrored	2.49	1.32
	Different	2.24	1.26
	Same	2.93	1.01





**Figure 9:** Mean scores for structure by high and low occlusion where mirrored and different-type items had reversed outcomes

To further explore this relationship, an exploratory factorial ANOVA was conducted to determine the relationship between the response patterns and occlusion conditions by structure type. The important thing to note here is that structure should not play a major part in the response patterning as we hypothesize that responses were mainly predicated on occlusion levels. The high occlusion items had a mean of 1.51 and an SD of 1.13, while low occlusion had a mean of 2.55 and an SD of 1.2 (see Table 7 for means and SDs broken down by structure type).

The assumptions for ANOVA were tested before proceeding with the analysis. The QQ plots were indicative of normally distributed data, while Levene's test based on the median was nonsignificant at  $p > 0.135$ . Using the same rationale as in the results section, a parametric ANOVA was conducted. The main effect of occlusion was statistically significant  $f(1,414) = 84.03, p < .001$ , with a partial eta squared of 0.169 indicating a large effect. The main effect of the structure was also significant  $f(2,414) = 4.1, p = 0.019$ , with a partial eta squared of 0.019, indicating a small effect. The interaction term of occlusion and structure was nonsignificant,  $p = 0.089$ .

Exploring the main effect of occlusion, the mean score for low occlusion is higher at 2.55, indicating that individuals answering questions containing a lower level of occlusion were more likely to say *yes*, further explaining why mirrored and different items had low means in the original dataset under low occlusion and higher means in the high occlusion. Pairwise comparisons were conducted to explore the effect of structure on response patterns. Post-hoc tests were conducted along with Bonferroni corrections. Findings indicated that there was a statistically significant difference between the same- and different-type items only. Different and mirrored items were in the same homogenous subset, while same-type items were in a higher and different subset. The finding that the same-type items would have the highest mean scores is not surprising considering these were the items where there was the least amount of manipulation, as “yes” was already the correct answer. The findings of the exploratory ANOVA provide support for the post-hoc conjecture that occlusion played a role in participant response patterns.

A possible reason for which we observe this responding pattern is that high levels of occlusion may instill a level of distrust within the participant, making them think that the assessment is trying to “trick” them. Hence, they answer “No” for those types of items. Future researchers should consider adding practice questions before presenting participants with the measure to ameliorate the level of distrust. Allowing participants to get a peek at the types of items they will encounter might give them the contextual knowledge they need to perform their best. This knowledge can show that the test creator is transparent about what is included in the measure. A limitation of this study was that practice questions were not offered. An image of how the items would be presented to

participants was shown, but it did not allow for participant interaction. It also does not give them any information on distinct types of items.

Lastly, to future researchers who want to use this measure and collect from online crowd-sourcing platforms, the researcher recommends several rules to follow when cleaning data before analysis. First, depending on the number of items included in the measure, check to see if the total time taken to complete was reasonable. For example, this measure had 32 items. Anyone who finishes this measure in under 2.5 minutes should be considered suspect. If they average 5-seconds per item, the measure should at least take them 2.6 minutes. Second, since it takes time to process the figures and move the mouse to answer and advance to the next screen, anyone who averages 2-seconds or less should be excluded, as they most likely responded carelessly. Lastly, response time mean and SD should be evaluated for each participant. Those with a deviation that is twice the mean or larger should be examined closely as this may indicate participants clicking through some items of the measure and letting the rest time out.

### **Incentivization and Motivation**

The second finding of this study was that there was a large amount of guessing and careless responding that resulted in low-quality data. Amazon's MTurk has been a data collection option for social scientists for around a decade. Thousands of studies and peer-reviewed articles have been published using data collected from this platform, with several researchers supporting the platform's ability to provide high-quality data (Hauser & Schwarz, 2016; Litman & Robinson, 2020; Litman et al., 2015), which was contradictory to the quality of data collected for this study. However, two recent articles discussed a new trend where there have been several reports of low-quality data being

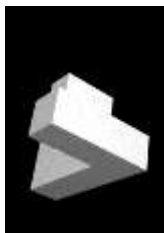
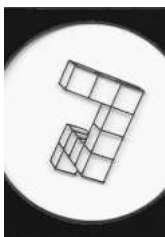
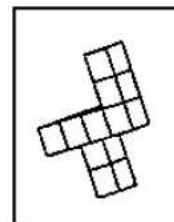
provided by MTurkers (Chandler et al., 2020; Kennedy et al., 2018). Several reasons include lack of proper motivation/incentivization, MTurker inattention, language comprehension troubles, effortless response, or fraudulent responses from individuals outside of the intended location for data collection. These reasons can have researchers that use MTurk as a data-collection tool questioning the accuracy of their data and subsequent conclusions.

To increase the probability of gathering quality data from MTurk, Litman and Robinson (2020) recommend several measures to take before data collection commencing. First, it is recommended to collect from participants that have a high HIT approval rate. Past behavior is the best predictor of future behavior. Therefore, those that have a high HIT approval rate should be selected. Highly rated MTurkers are more likely to provide high-quality data. This study set a limit to only allow MTurkers with a HIT approval rate of 95% or above to take the study. Second, provide clear instructions for the measure. The suggestion is straightforward—make sure everyone would understand what is asked of them and describe the study accurately so they know if this is the right HIT for them. Before collecting the data for this study, the committee and three peers reviewed the directions to ensure that clarity is maintained. Third, ensure that participants are paid a fair wage. Though data shows that pay does not impact data quality in multiple-choice survey HITs, this may not be true for other tasks (Litman et al., 2015). Since this study task was more cognitively demanding, we wanted to ensure participants were compensated fairly. The current recommended rate is 12 cents per minute (which is \$7.20 an hour—slightly below minimum wage). This study paid \$2 for a maximum possible time of 10 minutes (\$12.00 an hour, which is significantly higher than the

current minimum wage). Lastly, ensure the MTurk account has a good track record (pay on time, pay fairly, have a low HIT rejection rate, etc.). A bad reputation can deter high-quality candidates from taking the HIT. Candidates need to know they will be paid fairly and quickly for quality work. The account used to collect these data has a 92% HIT approval rate and has always paid above minimum wage.

This study followed all factors mentioned above: pay, good MTurk account reputation, clear instruction, and requesting from MTurkers with high HIT approvals and still ended up with a large percentage of low-quality data. A total of 180 HITs were requested. After cleaning procedures, only 70 HITs were deemed usable, which means that more than half of the HITs collected were low-quality data despite following best practices for MTurk data collection.

There are two potential reasons for this phenomenon. First, although they were given a significant pay incentive, the performance on the measure itself was not a factor. As long as they took the time to answer the questions, they got paid. The study offered no incentives for candidates to try to the best of their abilities. For cognitive tests in a business or academic setting, test-takers have something to gain or lose based on high or low test performance. The second hypothesis for the peculiar response pattern and low-quality data would be specifically related to the high level of occlusion. Compared to Shepard and Metzler and the Vandenberg and Kuse instruments, an extreme degree of occlusion was included in the present study (see Figure 10 for an example comparison).

**Present Study****Shepard-Metzler****Vandenburg-Kuse**

**Figure 10:** Comparison of occluded figures in this study to the two established MRTs

The high degree of occlusion of items in this study could contribute to the observed response patterns, suggesting that high occlusion appears to be related to guessing “no.” The high degree of occlusions raises the question — do very high levels of occlusion become a moderating variable in the motivation-performance relationship? Unfortunately, the current study did not have the data to test or make any conclusions in this regard and must leave this as an area of future research.

An additional safety measure that can be implemented might help with careless responding and allow for a more accurate measure of cognitive ability, including rewarding participants for quick and correct answers and penalizing them for incorrect responses and higher time. Based on the findings of this study and possible reasons for the findings, let us assume the high occlusion leads to guessing due to the lack of motivation, incentive, or the absence of a high-stake result (e.g., school admission, or job offer). Incentivizing correct and quick responses with money, and penalizing for long or incorrect responses, may provide the additional level of control necessary to collect quality data. Since response time is often an indicator of cognitive ability (quicker times indicative of higher levels of cognitive ability [processing speed]), introducing a scoring paradigm that includes participant response time might increase measurement accuracy

for spatial reasoning ability. Future research could determine whether this type of scoring procedure in a professional or academic context helps with careless responses.

Additionally, the scoring procedure, in-person proctoring, or remote (Zoom/WebEx) proctoring might reduce the number of careless responses and guessing, as participants may be more inclined to exert effort because they are being watched.

### **Angular Rotation when Considering AIG**

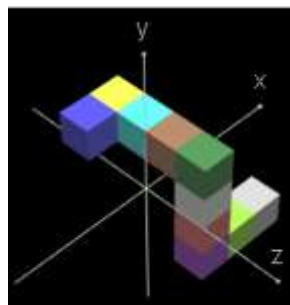
#### **Angular Rotation**

This study's angular rotation measurement method was intended to support that angular rotation in the items was independent of occlusion. Though it is established that angular rotation does not contribute to difficulty, it was important to show that the high occlusion items used here were not confounded with angular rotation for two reasons (Caissie et al., 2009). First, the degree of high occlusion present in this study has not been used before. Second, angular rotation was not held consistent when the stacks were originally being developed. A reasonable conclusion that angular rotation was likely not a confound was reached by establishing that the means and SDs of total degrees of rotation were not statistically significantly different between low and high occlusion groups.

After peer review, a limitation of the angular rotation measurement method was identified. Adding the absolute value of all three axes' rotation degrees does not provide an accurate total degree in rotation score. It is important to note that the R script is written to not provide a degree measurement by axes over 180. It automatically recodes anything over 180 as a proper negative rotation. For example, a 181-degree rotation would be recoded to -179 degrees. Although the current scoring method sufficed in the context of this study's items, applying it to AIG would inevitably fail. For example, if this method

of measuring angular rotation was used in an AIG framework, the resulting rotation score can be extremely misleading. For example, if an AIG framework generates a stack and rotates it 179 degrees on all three axes, the current method of measuring angular rotation would provide a rotation score of 537, indicating a very high degree of rotation. However, the figure is very close to being in its original position. Therefore, an alternative method was researched that provides a more accurate way of measuring triaxial rotation. This method of calculating a single score for angular rotation uses Euler Angles generated when a stack is rotated and transforms those angles into a 3X3 matrix. That matrix is then compared to a non-rotated stack matrix to determine the total degrees rotated.

A brief explanation is provided to show why Euler angles are important for rotation calculation for what they are and how they are derived. Each stack is rotated in ZYX succession, so the ZYX Euler angle formulas calculate the rotation. The order in which the rotations occur does indeed matter as different Euler formulas are used. Figure 11 shows how a cube stack is centered on an XYZ coordinate plane.

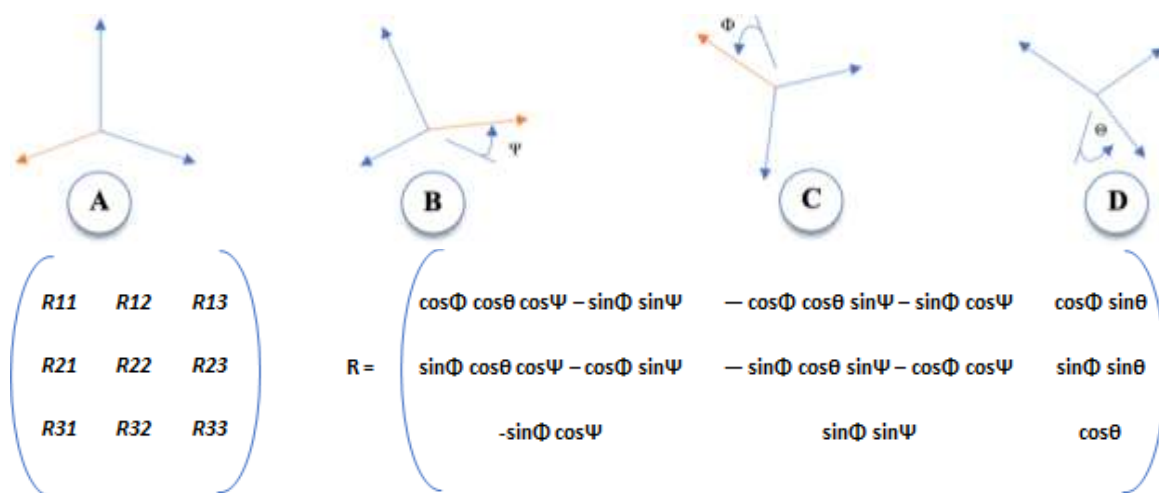


**Figure 11:** GIF/image of a stack rotation.

Breaking down Figure 11 into just the rotation is helpful when trying to conceptualize Euler angles (see Figure 12). To see what Euler angles are, we should think



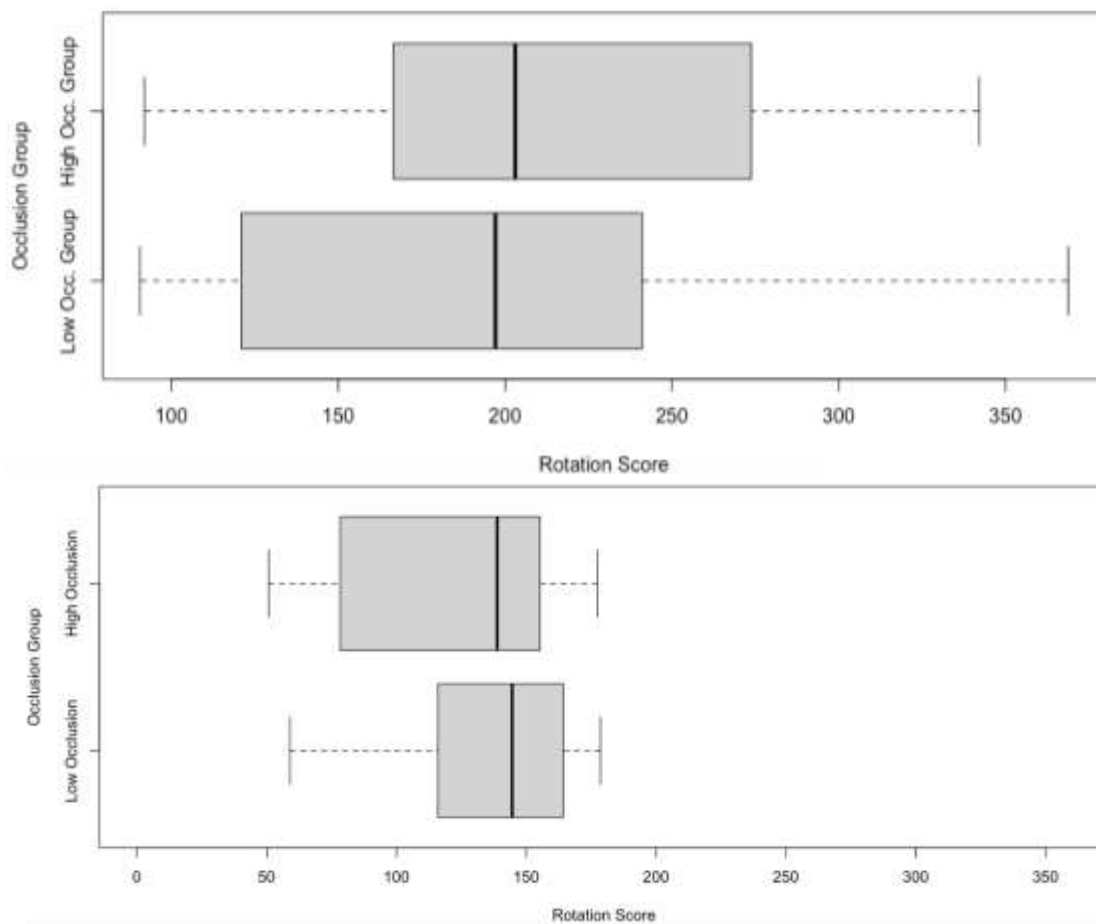
about the figure engaging in three successive rotations. In Figure 12, the three rotations are from A to B, B to C, and C to D. The orange arrow is the axis on which the figure is being rotated. The Greek characters below: A to B (Psi), B to C (Phi), and C to D (Theta) are Euler angles. The three Euler angles are transformed into a 3X3 matrix using the matrix formula below Figure 12, which is later used to calculate the total rotation of the cube stack.



**Figure 12:** Triaxial rotations for Euler angle calculation

The three numbers presented in Figure 2 on the row labeled “Rotation” are Euler angles calculated and recorded. At the same time, the stacks were being physically rotated in the development stage of the items in a box similar to what is shown in Figure 11. The package RGL containing the function Rotate3D (see GitHub for the package script) is responsible for calculating the Euler angles after the stack has been created. Then, using matrix algebra and the RGLtoLATTICE function, we can compare the total angular rotation of each stack to an unrotated stack (i.e., 1s in the diagonal) and arrive at the value of total degrees rotated.

The script was tested using various examples. Two noteworthy examples show the effectiveness of the improved process. The first test was to create three stacks and rotate the first stack by 45 degrees on the x-axis, the second stack by 45 degrees on the y-axis, and the third stack by 45 degrees on the z-axis. Calculations showed all three stacks having total rotations of 45 degrees. The second test was to rotate a stack on all three axes by 179 degrees, and the total degrees of rotation was 1.703. Recall the example of why the original method would not work in an AIG framework. Rotating 179 degrees on all axes gives us a total rotation of 537 degrees. The problem is that although the shape was rotated 537 degrees, the actual position of the shape is equivalent to slightly rotating it one degree on each axis. The Euler angle method of measuring total rotation arrives at a total rotation of 1.703 degrees, a more practical and accurate way of looking at total rotation concerning the final position being virtually the same. These examples tested the extremes—little rotation to extreme rotation—and yet showed expected results based on the original positioning of the stack. The bottom plot in Figure 13 shows the recalculated angular rotation scores for low and high occlusion groups.



**Figure 13:** Box and whisker plot comparisons between the first (top plot) and second (bottom plot) calculation methods for angular rotation scores

There are two noticeable differences in the results from the two methods (see Figure 13): range of angular rotations and means for high and low occlusion between methods. For the first method, the scores ranged from 90.53 to 368.70, while the more accurate Euler angle method ranged from 50.77 to 179. Notice how the first method has a rotation score above 360 while the second does not exceed 179.

The difference between the mean rotation scores for low and high occlusion groups when calculated using the matrix method shows no statistically significant difference  $t(22) = 0.72$ ,  $p = 0.48$ , supporting that angular rotation was consistent among

both groups, not a likely confound. Nevertheless, future researchers are encouraged to use this method when developing MRTs, especially in an AIG context when studying the link between occlusion and angular rotation.

### **Conclusion**

Technology has made it easier for organizations and test providers to mass administer testing content quickly and remotely and comes with a string of test security concerns, giving test administrators a valid reason to doubt the validity of the test scores. AIG measures are great tools for combating these test security concerns.

A key step in building/developing AIG measures is to successfully identify radicals, the item features that contribute to the difficulty of test items. This study takes a special interest in occlusion and structure type within the MRT space to build the groundwork for the future development of MRT using AIG. The research ideas that were the basis of the current study were used to develop a computerized version of the SMMRT programmed in R to collect data on occlusion effects on MRT test items. In sum, the study found no support for the hypotheses but discovered a peculiar response pattern that has not been discussed within the MRT space based on the literature review. It is hypothesized that this response pattern is related to the levels of occlusion; however, a limitation of this study is that it dichotomizes occlusion to high and low. As stated above, the high occlusion items used here are much more occluded than Sheppard-Metzler and Vandenburg-Kuse, so this responding pattern might only occur when more extreme high occlusion is present. Future researchers studying this should consider using a continuous scale of occlusion to determine if this does indeed happen on the end of high occlusion and low occlusion items.

While this study does not provide unequivocal support for using occlusion as a radical in an AIG methodology, it highlights many opportunities for future researchers to pursue several lines of research and continue building an AIG version of the SMMRT.

## REFERENCES

- Arendasy, M., and Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes assessment. *Learning and Individual Differences, 22*, 112–117. doi: 10.1016/j.lindif.2011.11.005
- Arendasy, M. E., and Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*(3), 181-192. doi: 10.1016/j.intell.2013.02.004
- Block, N. (1993). Book review of Dennett's consciousness explained. *Journal of Philosophy, 90*, 93-181.
- Caissie, A., Vigneau, F., and Bors, D. (2009). What does the mental rotation test measure? An analysis of item difficulty, and item characteristics. *Open Psychology Journal, 2*, 94-102. doi: 10.2174/1874350100902010094
- Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness and affirmative action. *Journal of Vocational Behaviour, 49*, 122-158.  
doi:10.1006/jvbe.1996.0038
- Carpenter, P. A., and Just, M. A. (1978). Eye fixations during mental rotation. In J. W. Senders, D. F. Fisher, and R. A. Monty (Eds.), *Eye movements and the higher psychological junctions* (pp. 115-133). Erlbaum.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

- Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, 31, 176–199. doi: 10.1037/h0059043
- Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, 38, 592.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Houghton Mifflin.
- Cattell, R. B., and Cattell, A. K. S. (1963). *Culture Fair Intelligence Test*. Institute for Personality and Ability Testing.
- Chandler J., Sisso I., Shapiro D. (2020). Participant carelessness and fraud: consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, 129: 49–55.
- Chernyshenko, O. S., and Stark, S. (2015). Mobile psychological assessment. In F. Drasgow, Ed., *Technology and Testing: Improving Educational and Psychological Measurement* (Vol. 2). Wiley-Blackwell.
- Cook, L. L., and Eignor, D. R. (1991). NCME Instructional module: IRT equating methods. *Educational Measurement, Issues and Practice*, 10, 37–45.
- Cukusic, M., Garaca Z., and Jadric, M. (2014). Online self-assessment and students' success in higher education institutions. *Computers and Education* 72, 100–109. doi:10.1016/j.compedu.2013.10.018.
- Drasgow, F., Nye, C., Guo, J., and Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology*, 2(1), 46-48. doi:10.1111/j.1754-9434.2008.01106.x
- Field, A. P. (2009). *Discovering statistics using SPSS: And sex and drugs and rock 'n' roll*. SAGE Publications.

- Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Sage.
- Foster, D. F. (2010). Worldwide testing and test security issues: Ethical challenges and solutions. *Ethics and behavior*, 20(3-4), 207-228.  
doi: 10.1080/10508421003798943
- Furnham, A. (2008). *Personality and intelligence at work: Exploring and explaining individual differences at work*. London: Routledge.
- Geerlings, H., Glas, C., and van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76, 337-359. doi: 10.1007/S11336-011-9204-X
- Georgopoulos, A., Lurito, J., Petrides, M., Schwartz., A., and Massey, J. (1989). Mental rotation of the neuronal population vector. *Science*. Jan 13;243(4888):234-6. doi: 10.1126/science.2911737.
- Gierl, M. J., and Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. Routledge.
- Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, 15, 25–46. doi: 10.1207/S15327043HUP1501&02\_03
- Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology*, 86(1), 174-199. doi: 10.1037/0022-3514.86.1.174
- Guilford, J. P. (1967). *The Nature of Human Intelligence*. McGraw-Hill.
- Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement*, 48, 1–4. doi:10.1177/ 001316448804800102



- Hausdorf, P. A., LeBlanc, M. M., and Chawla, A. (2003). Cognitive ability testing and employment selection: Does test content relate to adverse impact? *Applied H.R.M. Research*, 7(1-2), 41-48.
- Hauser, D., Paolacci, G., and Chandler, J. (2018). *Common Concerns with MTurk as a Participant Pool: Evidence and Solutions*. <https://doi.org/10.31234/osf.io/uq45c>
- Hauser, D. and Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavioral Research* 48, 400–407. <https://doi.org/10.3758/s13428-015-0578-z>
- Hertzog, C., and Rypma, B. (1991). Age differences in components of mental-rotation task performance. *Bulletin of the Psychonomic Society*, 29(3), 209-212.
- Hochberg, J., and Gellman, L. (1977). The effect of landmark features on mental rotation times. *Memory and Cognition* 5, 23–26. <https://doi.org/10.3758/BF03209187>
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, and R. W. Woodcock, *Woodcock–Johnson technical manual* (pp. 197–232).
- Horn, J. L., and Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57(5), 253–270. <https://doi.org/10.1037/h0023816>
- Hulsheger, U., Maier, G., and Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany. *International Journal of Selection and Assessment*, 15, 3–18. doi: 10.1111/j.1468-2389.2007.00363.x

- Irvine, S. (2002). The foundations of item generation for mass testing. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development* (pp.3-32). Erlbaum.
- Johnson, A. M. (1990). Speed of mental rotation as a function of problem-solving strategies. *Perceptual and Motor Skills*, 71(3), 803-806.
- Jones, B., and Anuza, T. (1982). *Effects of sex, handedness, stimulus and visual field on 'mental rotation.'* *Cortex*, 18, 501-14.
- Just, M., and Carpenter, P. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 12, 21-31.
- Just, M. A., and Carpenter, P. A. (1975). Eye fixations and cognitive processes. *Complex Information Processing Working Papers, No. 29*. Carnegie-Mellon University.
- Karim, M., Kaminsky, S., and Behrend, T (2014). Cheating, reactions, and performance in remotely proctored testing: An Exploratory experimental study. *Journal of Business Psychology*, 29, 555–572, doi.org/10.1007/s10869-014-9343-z
- Kennedy, R., Clifford, S., and Burleigh, T. (2018). The shape of and solutions to the mturk quality crisis. Available at SSRN: <https://ssrn.com/abstract=3272468> or <http://dx.doi.org/10.2139/ssrn.3272468>
- Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the SIGCHI conference on human factors in computing systems*, 453-456.
- Koshino, H., Carpenter, P. A., Minshew, N. J., Cherkassky, V. L., Keller, T. A., and Just, M. A. (2005). Functional connectivity in an fMRI working memory task in high-functioning autism. *NeuroImage*, 24, 810–821.

- Kozlowski S., and Bell, B. (2012). Work groups and teams in organizations. *Handbook of Industrial and Organizational Psychology, 12*,  
doi.org/10.1002/9781118133880.hop212017
- Lai H., Gierl M. J., Byrne B. E., Spielman A., and Waldschmidt D. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education, 80*, 339-347. [[Google Scholar](#)]
- Litman L., and Robinson J. (In Press, 2020). *Conducting online research on amazon mechanical turk and beyond*. Sage Publications.
- Litman L., Robinson J., and Rosenzweig C. (2014). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavioral Research Methods, 47*(2):519-528.  
doi:10.3758/s13428-014-0483-x
- Marks, D. F. (1999). Consciousness, mental imagery and action. *British Journal of Psychology, 90*, 567-585.
- McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf–Gc framework. In D. P. Flanagan, J. L. Genshaft, and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). Guilford.
- Monahan, J. S., Harke, M. A., and Shelley, J. R. (2008). Computerizing the mental rotations test: are gender differences maintained? *Behavior Research Methods, 40*, 422–427. doi:10.3758/BRM.40.2.422.
- Neisser, U. (1967). *Cognitive psychology*. Appleton Century Crofts.

- Ones, D., Viswesvaran, C., and Dilchert, S. (2006). Cognitive ability in selection decisions. In D. Wilhemi and R. Engle (Eds.), *Understanding and measuring intelligence*. Sage.
- O'Boyle, M., Cunnington, R., Silk, T., Vaughan, D., Jackson, G., Syngeniotes, A., (2005). Mathematically gifted male adolescents activate a unique brain network during mental rotation. *Cognitive Brain Research*, 25(2), 583-587.
- Oostra K. M., Vereecke A., Jones K., Vanderstraeten G., and Vingerhoets G. (2012). Motor imagery ability in patients with traumatic brain injury. *Arch Phys Med Rehabil*. 2012 May;93(5):828-33. doi: 10.1016/j.apmr.2011.11.018.
- Ployhart, R. E., and Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61(1), 153-172.
- Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. SpringerLink.  
doi: [https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-79948-3\\_1068](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-79948-3_1068)
- Raven, J., and Court, J. (1989). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research supplement no. 4: Additional national and American norms, and summaries of normative, reliability, and validity studies*. The Psychological Corporation.
- Reeves, T. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research*. 23(1), 101-11.  
doi: 10.2190/FYMQ-78FA-WMTX-J06C

- Richter, W., Somorjai, R., Summers, R., Jarmasz, M., Menon, R., and Gati, J. (2000). Motor area activity during mental rotation studied by time-resolved single-trial fMRI. *Journal of Cognitive Neuroscience*, *12*(2), 310-320.
- Schmidt, F. L., and Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162-173. doi: 10.1037/0022-3514.86.1.162
- Sheets, T. L. (2020). Shepard-Metzler Automated Item Generation project.  
<https://github.com/katyem/smaig/>
- Shepard, R., and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*,701-703.
- Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–293.  
<https://doi.org/10.2307/1412107>
- Spearman, C. (1927). *The nature of intelligence and the principles of cognition*. Macmillan and Co.
- Steiger, J. H., and Yuille, J. C. (1983). Long-term memory and mental rotation. *Carbadian Journal of Psychology*, *37*, 367-389.
- Stevens, J. (1996). Applied multivariate statistics for the social sciences, 4th ed. Erlbaum.
- Tippins, N. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*. *2*, 551-582.  
<https://doi.org/10.1146/annurev-orgpsych-031413-091317>

- van der Maas, H., Kan, K., and Borsboom, D. (2014). Intelligence is what the intelligence test measures: Seriously. *Journal of Intelligence*, 2(1), 12-15. doi: 10.3390/jintelligence2010012
- Vandenberg, S. G. (1975). Sources of variance in performance on spatial tests. In I. Eliot and N. I. Salkind (Eds.), *Children 50 spatial development* (pp. 57-66). Charles C. Thomas.
- Vandenberg S. and Kuse A. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Percept Mot Skills*. 47(2):599-604. doi: 10.2466/pms.1978.47.2.599. PMID: 724398.
- Voyer, D. and Hou, J. (2006). Type of items and the magnitude of gender differences on the mental rotations test. *Canadian Journal of Experimental Psychology*, 60, 91-100.
- Wilcox, R. R. (2005). New methods for comparing groups: Strategies for increasing the probability of detecting true differences. current directions in *Psychological Science*, 14(5), 272–275. <https://doi.org/10.1111/j.0963-7214.2005.00379.x>
- Wobbrock, J., Findlater, L., Gergle D., and Higgins, J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '11). ACM Press, pp. 143-146.
- Zenisky, A. L., and Sireci, S. G. (2013) Innovative items to measure higher-order thinking: Development and validity considerations. *Presentation at the annual meeting of the National Council of Measurement in Education*, San Francisco, CA

## **APPENDIX A**

### **DEMOGRAPHIC ITEMS**

### Demographic Items

All of the demographic items that participants were asked along with all possible response options.

- What is your gender?
  - Male, Female, Other (please specify)
- What is your age?
  - 18-80
- With which race or ethnicity do you most identify?
  - White, Hispanic or Latino, Black or AA, Asian, Pacific Islander, Native Hawaiian, Native American, Indian American (Indian subcontinent), Other (Please Specify)
- Select your highest completed level of education
  - No Schooling completed to Doctorate Degree
- Please indicate your level of familiarity when it comes to using computers or internet-enabled devices.
  - Never used one before (0) to Expert (100)

Please refer to Appendix C for the reference Images. MRT Question Matrix in Appendix D



**APPENDIX B**

**HUMAN USE APPROVAL LETTER**



Office of Sponsored Projects

## EXEMPTION MEMORANDUM

TO: Mr. Swadeep Patel and Dr. Tilman Sheets

FROM: Dr. Richard Kordal, Director of Intellectual Properties  
 rkordal@latech.edu

SUBJECT: HUMAN USE COMMITTEE REVIEW

DATE: March 15, 2021

TITLE: "Exploring the Effect of Low & High Levels of Occlusion on a  
 Computerized Mental Rotation Test:  
 Implications for Automatic Item Generation"

NUMBER: HUC 21-073

According to the Code of Federal Regulations Title 45 Part 46, your research protocol is determined to be exempt from full review under the following exemption category(s):  
 46.104 (d) (2) (i) (ii).

(d) Except as described in paragraph (a) of this section, the following categories of human subjects research are exempt from this policy:

(2) Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) if at least one of the following criteria is met:

(i) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects;








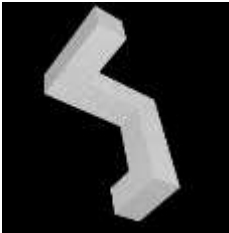













(ii) Any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation; or

Thank you for submitting your Human Use Proposal to Louisiana Tech's Institutional Review Board.







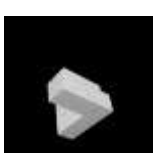
## **APPENDIX C**

### **CUBE STACKS, OCCLUSION INDICATORS, AND SCORES**

Record of All Cube Stacks and Occlusion Indicators and Scores (Stacks 1-3). Each image listed under columns C-E are individual items.

Original Cube Stacks (A)	Occlusion Level (B)	Mirrored Stack (C)	Different Stack (D)	Same Stack (E)	
<p>Cube Stack 1</p>  <p>Occlusion Score: .576</p>	Low	 OS: .585	 OS: .589	 OS: .558	
	High	 OS: .645	 OS: .654	 OS: .665	
	<p>Cube Stack 2</p>  <p>Occlusion Score: .565</p>	Low	 OS: .563	 OS: .595	 OS: .548
		High	 OS: .626	 OS: .734	 OS: .644
<p>Cube Stack 3</p>  <p>Occlusion Score: .562</p>		Low	 OS: .561	 OS: .516	 OS: .580
		High	 OS: .764	 OS: .638	 OS: .762

## C: Continued (Stack 4)

Original Cube Stacks	Occlusion Level	Mirrored Stack	Different Stack	Same Stack
<p data-bbox="298 373 472 405">Cube Stack 4</p>  <p data-bbox="298 640 521 709">Occlusion Score: .568</p>	Low	 OS: .467	 OS: .506	 OS: .524
	High	 OS: .715	 OS: .748	 OS: .734

## **APPENDIX D**

### **MATRICES**

This table contains the matrices used to develop the cube stacks for this study. R script is available on [GitHub](#) for researchers interested in recreating these items.

Stack ID	x-Axis	y-Axis	z-Axis	Leg Count
1	-1	0	0	1
1	-2	0	0	1
1	-2	0	-1	2
1	-2	0	-2	2
1	-2	0	-3	2
1	-2	1	-3	3
1	-2	2	-3	3
1	-2	3	-3	3
1	-3	3	-3	4
1	-4	3	-3	4
Rotations	-19.8637	59.03442	-0.63226	1
3	1	0	0	1
3	2	0	0	1
3	3	0	0	1
3	3	0	-1	2
3	3	0	-2	2
3	3	0	-3	2
3	3	1	-3	3
3	3	2	-3	3
3	3	3	-3	4
3	4	3	-3	4
Rotations	-35.8562	-52.9005	-167.242	1
4	0	0	1	1
4	0	0	2	1
4	0	1	2	2
4	0	2	2	2
4	0	3	2	2
4	0	4	2	2
4	0	4	3	3
4	0	4	4	3
4	-1	4	4	4
4	-2	4	4	4
Rotations	101.1561	-31.3371	135.6969	1
7	0	1	0	1
7	0	2	0	1
7	0	3	0	1
7	0	3	1	2
7	0	3	2	2
7	0	3	3	2
7	1	3	3	3
7	2	3	3	3
7	2	4	3	4
7	2	5	3	4

Rotations	147.9856	-13.4158	-177.784	1
8	-1	0	0	1
8	-2	0	0	1
8	-2	0	-1	2
8	-2	0	-2	2
8	-2	0	-3	2
8	-2	1	-3	3
8	-2	2	-3	3
8	-2	3	-3	3
8	-3	3	-3	4
8	-4	3	-3	4
Rotations	-49.883	23.9963	-42.8735	1
9	-1	0	0	1
9	-2	0	0	1
9	-2	0	-1	2
9	-2	0	-2	2
9	-2	0	-3	2
9	-2	1	-3	3
9	-2	2	-3	3
9	-2	3	-3	3
9	-3	3	-3	4
9	-4	3	-3	4
Rotations	-51.828	50.01707	-11.7341	1
10	1	0	0	1
10	2	0	0	1
10	2	0	1	2
10	2	0	2	2
10	2	0	3	2
10	2	-1	3	3
10	2	-2	3	3
10	2	-3	3	3
10	3	-3	3	4
10	4	-3	3	4
Rotations	59.12584	-8.10798	141.6815	1
11	1	0	0	1
11	2	0	0	1
11	2	0	1	2
11	2	0	2	2
11	2	0	3	2
11	2	-1	3	3
11	2	-2	3	3
11	2	-3	3	3
11	3	-3	3	4
11	4	-3	3	4
Rotations	41.82184	-48.6287	158.6188	1
12	0	1	0	1



12	0	2	0	1
12	0	3	0	1
12	0	3	1	2
12	0	3	2	2
12	0	3	3	2
12	1	3	3	3
12	2	3	3	3
12	2	4	3	4
12	2	5	3	4
Rotations	163.4044	-40.4953	-50.2637	1
13	0	1	0	1
13	0	2	0	1
13	0	3	0	1
13	0	3	1	2
13	0	3	2	2
13	0	3	3	2
13	1	3	3	3
13	2	3	3	3
13	2	4	3	4
13	2	5	3	4
Rotations	92.50622	-35.6986	-70.3592	1
20	1	0	0	1
20	2	0	0	1
20	3	0	0	1
20	3	0	-1	2
20	3	0	-2	2
20	3	0	-3	2
20	3	1	-3	3
20	3	2	-3	3
20	3	3	-3	4
20	4	3	-3	4
Rotations	-44.448	20.99398	25.08327	1
21	1	0	0	1
21	2	0	0	1
21	3	0	0	1
21	3	0	-1	2
21	3	0	-2	2
21	3	0	-3	2
21	3	1	-3	3
21	3	2	-3	3
21	3	3	-3	4
21	4	3	-3	4
Rotations	-162.661	-30.1573	82.13157	1
22	-1	0	0	1
22	-2	0	0	1
22	-3	0	0	1

22	-3	0	1	2
22	-3	0	2	2
22	-3	0	3	2
22	-3	-1	3	3
22	-3	-2	3	3
22	-3	-3	3	4
22	-4	-3	3	4
Rotations	27.81161	-21.0706	179.1065	1
23	-1	0	0	1
23	-2	0	0	1
23	-3	0	0	1
23	-3	0	1	2
23	-3	0	2	2
23	-3	0	3	2
23	-3	-1	3	3
23	-3	-2	3	3
23	-3	-3	3	4
23	-4	-3	3	4
Rotations	-7.04906	24.97484	-152.99	1
24	0	0	1	1
24	0	0	2	1
24	0	1	2	2
24	0	2	2	2
24	0	3	2	2
24	0	4	2	2
24	0	4	3	3
24	0	4	4	3
24	-1	4	4	4
24	-2	4	4	4
Rotations	123.9501	47.63705	133.0603	1
25	0	0	1	1
25	0	0	2	1
25	0	1	2	2
25	0	2	2	2
25	0	3	2	2
25	0	4	2	2
25	0	4	3	3
25	0	4	4	3
25	-1	4	4	4
25	-2	4	4	4
Rotations	145.1434	49.46879	147.4153	1
26	0	0	1	1
26	0	0	2	1
26	0	1	2	2
26	0	2	2	2
26	0	3	2	2

26	0	4	2	2
26	0	4	3	3
26	0	4	4	3
26	-1	4	4	4
26	-2	4	4	4
Rotations	-118.48	-29.5563	-29.8454	1
27	0	0	1	1
27	0	0	2	1
27	0	1	2	2
27	0	2	2	2
27	0	3	2	2
27	0	4	2	2
27	0	4	3	3
27	0	4	4	3
27	-1	4	4	4
27	-2	4	4	4
Rotations	-149.553	-38.6891	-19.2932	1
28	0	0	-1	1
28	0	0	-2	1
28	0	-1	-2	2
28	0	-2	-2	2
28	0	-3	-2	2
28	0	-4	-2	2
28	0	-4	-3	3
28	0	-4	-4	3
28	1	-4	-4	4
28	2	-4	-4	4
Rotations	61.28413	-32.7143	-31.1621	1
29	0	0	-1	1
29	0	0	-2	1
29	0	-1	-2	2
29	0	-2	-2	2
29	0	-3	-2	2
29	0	-4	-2	2
29	0	-4	-3	3
29	0	-4	-4	3
29	1	-4	-4	4
29	2	-4	-4	4
Rotations	66.68644	-25.4721	-61.2755	1
30	1	0	0	1
30	2	0	0	1
30	3	0	0	1
30	3	0	-1	2
30	3	0	-2	2
30	3	0	-3	2
30	3	1	-3	3

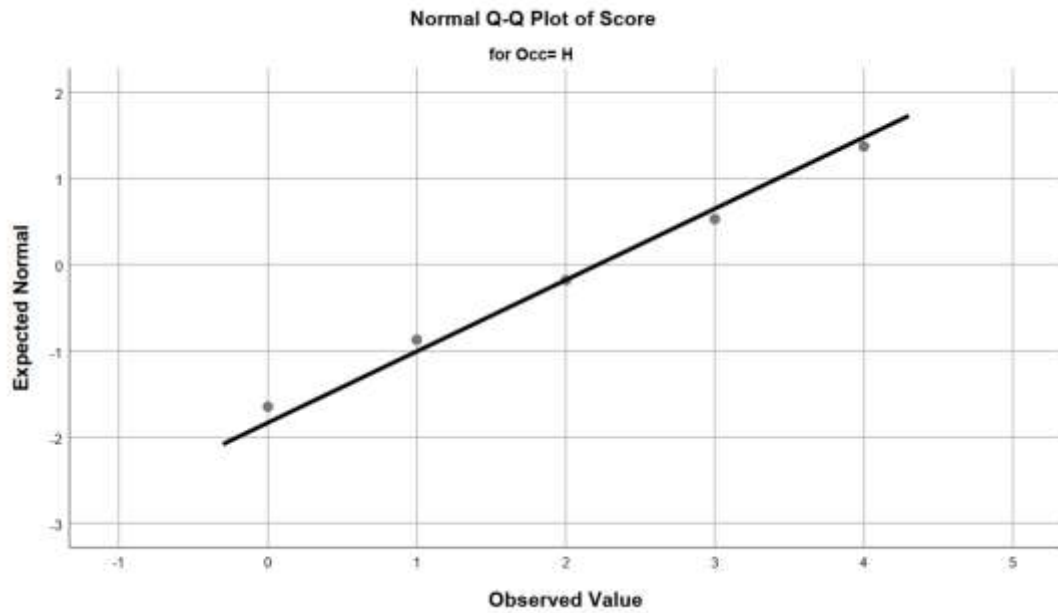
30	3	2	-3	3
30	3	3	-3	4
30	4	3	-3	4
Rotations	37.78183	-3.62531	-143.868	1
31	1	0	0	1
31	2	0	0	1
31	3	0	0	1
31	3	0	-1	2
31	3	0	-2	2
31	3	0	-3	2
31	3	1	-3	3
31	3	2	-3	3
31	3	3	-3	4
31	4	3	-3	4
Rotations	121.0059	-24.1214	-131.025	1
38	0	1	0	1
38	0	2	0	1
38	0	3	0	1
38	0	3	1	2
38	0	3	2	2
38	0	3	3	2
38	1	3	3	3
38	2	3	3	3
38	2	4	3	4
38	2	5	3	4
Rotations	50.71451	6.62831	49.11779	1.1025
39	0	1	0	1
39	0	2	0	1
39	0	3	0	1
39	0	3	1	2
39	0	3	2	2
39	0	3	3	2
39	1	3	3	3
39	2	3	3	3
39	2	4	3	4
39	2	5	3	4
Rotations	139.7076	1.4755	38.41332	1.1025
40	0	-1	0	1
40	0	-2	0	1
40	0	-3	0	1
40	0	-3	-1	2
40	0	-3	-2	2
40	0	-3	-3	2
40	-1	-3	-3	3
40	-2	-3	-3	3
40	-2	-4	-3	4

40	-2	-5	-3	4
Rotations	141.5955	47.94441	-179.162	1.1025
41	0	-1	0	1
41	0	-2	0	1
41	0	-3	0	1
41	0	-3	-1	2
41	0	-3	-2	2
41	0	-3	-3	2
41	-1	-3	-3	3
41	-2	-3	-3	3
41	-2	-4	-3	4
41	-2	-5	-3	4
Rotations	16.33831	41.01675	34.58241	1.1025
42	1	0	0	1
42	2	0	0	1
42	3	0	0	1
42	4	0	0	1
42	4	0	1	2
42	4	0	2	2
42	4	1	2	3
42	4	2	2	3
42	5	2	2	4
42	6	2	2	4
Rotations	-129.346	-40.1401	40.18424	1
43	1	0	0	1
43	2	0	0	1
43	3	0	0	1
43	4	0	0	1
43	4	0	1	2
43	4	0	2	2
43	4	1	2	3
43	4	2	2	3
43	5	2	2	4
43	6	2	2	4
Rotations	-120.327	-57.8193	94.53301	1

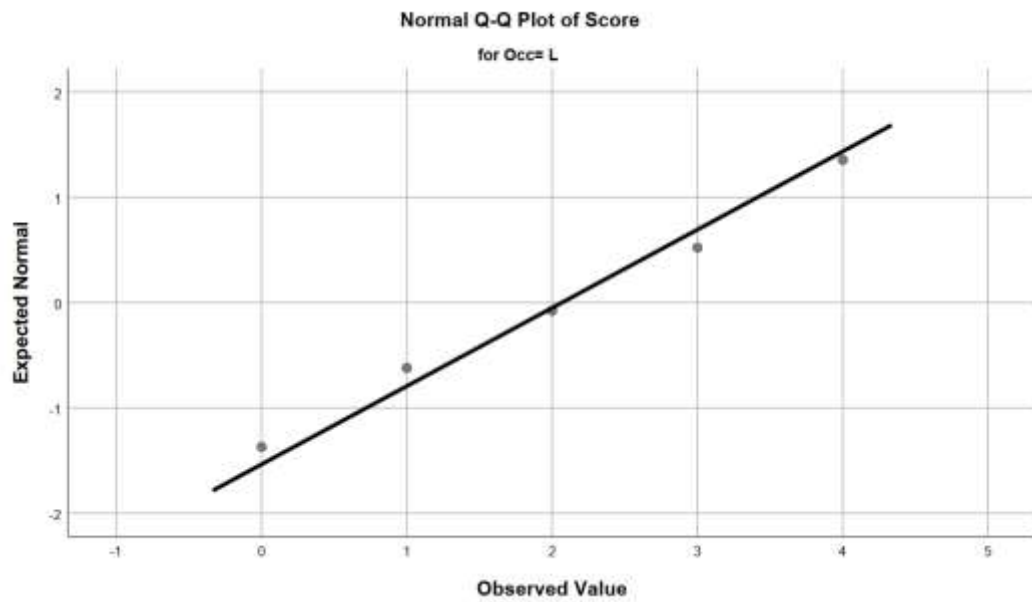
## **APPENDIX E**

### **QQ PLOT FOR HIGH OCCLUSION**

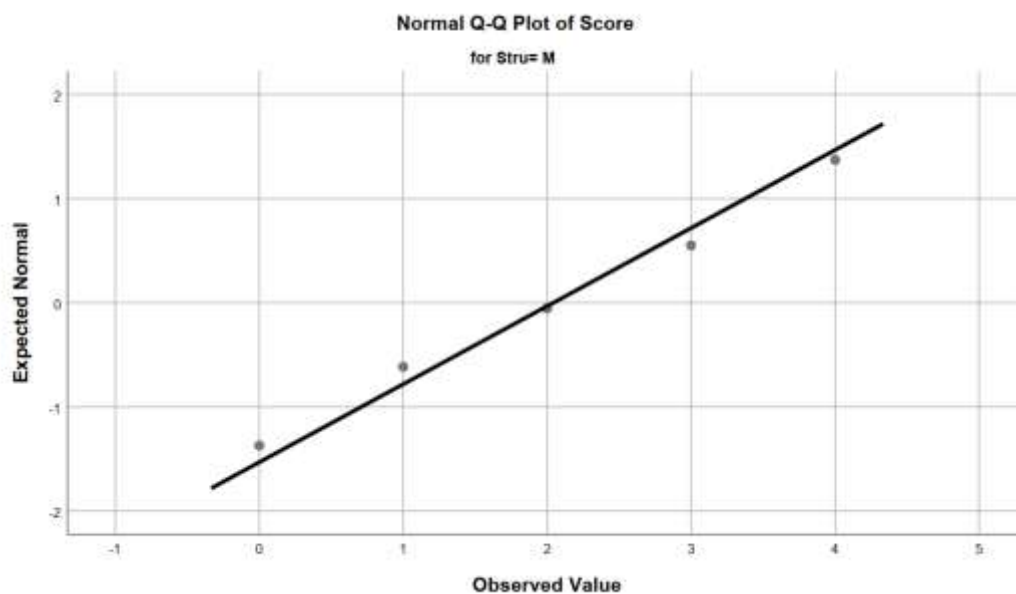
### QQ Plot for High Occlusion



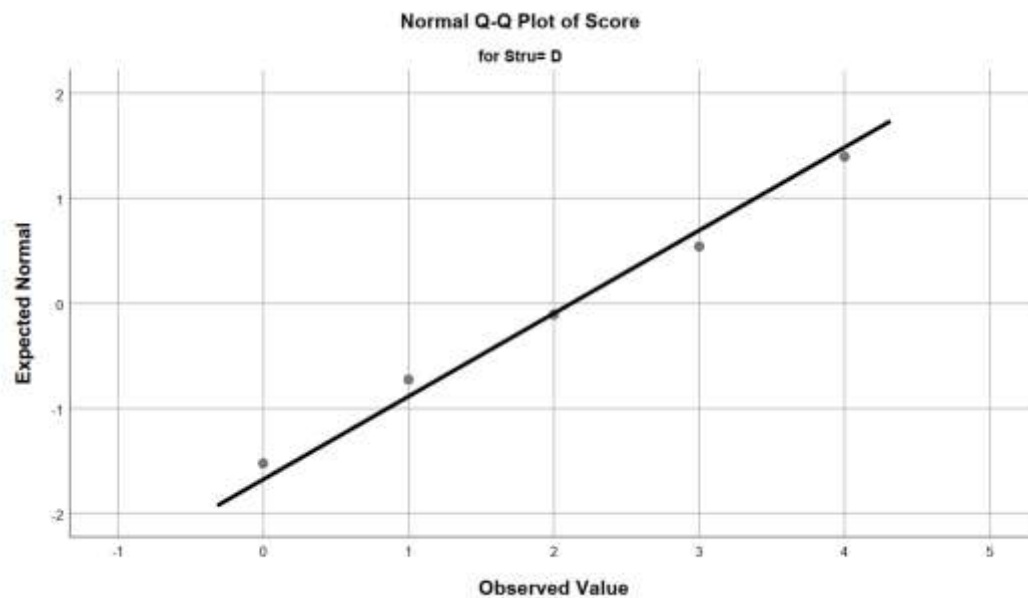
### QQ Plot for Low Occlusion



### QQ Plot for Mirrored Structure



### QQ Plot for Different Structure





## QQ Plot for Same Structure

