Evaluating Shadowspect as a Potential Measure of Spatial Reasoning

by

Melat R. Anteneh

B.S. Computation and Cognition

Massachusetts Institute of Technology, 2020

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES AND DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTATION AND COGNITION
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author: _____

Melat Anteneh
Department of Brain and Cognitive Sciences
Department of Electrical Engineering and Computer Science May 14, 2021

Certified by:_____

Eric Klopfer
Professor of Comparative and Media Studies
Thesis Supervisor

Accepted by: _____

Michale Fee
Professor of Brain and Cognitive Sciences
Associate Department Head of Education

ABSTRACT

Spatial reasoning allows individuals to conceive and manipulate mental representations of objects in space and is an essential process in countless daily activities (Clements & Battista, 1992). The online geometric puzzle game Shadowspect was created as a tool to evaluate players' spatial reasoning skills. The goal of this project was to evaluate Shadowspect's potential as a spatial reasoning assessment by comparing performance on the game to that on Ramful, Lowrie, and Logan's (2016) validated Spatial Reasoning Instrument. Shadowspect performance was strongly correlated to performance on the Spatial Reasoning Instrument, particularly when measured as a function of average solve time, i.e., the average time spent solving a puzzle ($r = -0.579,\ p < .001$) and total number of levels completed ($r = 0.705,\ p < .001$). The results of this study indicate that Shadowspect has the capability to serve as a measure of spatial reasoning.

INTRODUCTION

Spatial reasoning is the process of mentally representing, manipulating, and transforming objects or their subparts, both individually and in the context of their environments (Burnet & Lane, 1980; Clements & Battista, 1992). Spatial reasoning is composed of three interconnected sub-processes: mental rotation, spatial orientation, and spatial visualization. Together, these components allow individuals to create complex and dynamic multidimensional mental representations (Pittalis & Christou, 2010; Lohman, 1979). Mental rotation refers to the process of accurately creating internal representations of 2D and 3D objects from various perspectives (Shepard & Metzler, 1971). Spatial orientation is the process of manipulating and comprehending an object's position and heading in its environment and relating this information to an egocentric reference frame (Hegarty & Waller, 2004; Velez, Silver, & Tremaine, 2005). Spatial visualization is the ability to conceptualize the effects of imagined actions on an object's state, such as folding or unfolding, assembly and disassembly, turning, and moving (Burnet & Lane, 1980; Gorska & Sorby, 2008; McGee, 1979; Velez, Silver, & Tremaine, 2005). Unlike with mental rotation, spatial visualization may involve rotating an object so the representation of the object's transformed position or orientation is in relation to another, fixed object or point (Velez, Silver, & Tremaine, 2005).

Greater levels of spatial reasoning has been found to positively correlate with a number of favorable outcomes, including mathematical ability throughout childhood and adolescence, particularly in geometry (Clements & Battista, 1992; Pittalis & Christou, 2010; Verdine, Golinkoff, Hirsh-Pasek, & Newcombe, 2017). These positive outcomes have reportedly carried over into performance in science, technology, engineering, and math (STEM)-related career fields in adulthood (Hsi, Linn, & Bell, 1997; Lubinski, 2010; Wai, Lubinski, & Benbow, 2009).

Despite being a potential highly predictive measure of positive outcomes, spatial reasoning ability is not a fixed trait. Notably, spatial reasoning intervention has been found to successfully improve both spatial reasoning ability (Ben-Chaim, Lappan, & Houang, 1988; Clements & Battista, 1992; Ehrlich, Levine, & Goldin-Meadow, 2006; Eraso, 2007; Septia & Prahmana, 2018) and STEM achievement (Cheng & Mix, 2014; Lowrie, Logan, & Hegarty, 2019; Stieff & Uttal, 2015), especially in children and young adults. Playing games has been found to not only correlate with higher spatial reasoning ability but to be an effective method of improving spatial reasoning ability in general (Corradini, 2011), as well as mental rotation (Cherney, 2008; De Lisi & Wolford, 2002), spatial orientation (McClurg & Chaillé, 1987), and spatial visualization (Dorval & Pepin, 1986) ability specifically. Game-based intervention was found to be particularly less effective for older individuals (Gagnon, 1985). The favorable outcomes associated with higher spatial reasoning ability combined with the decrease in intervention effectiveness associated with age highlights the importance of detecting spatial reasoning skills early, particularly in school age children and young adults, to identify individuals who may need spatial reasoning intervention.

Spatial Reasoning & Gender

Considering the benefits of spatial reasoning ability for future outcomes, it is important to consider demographic factors that can potentially influence an individuals' spatial reasoning ability. For example, men have historically performed better than women on spatial assessments, including measures of spatial reasoning (Halpern 2013; Linn & Petersen, 1985; Maccoby & Jacklin, 1974). While the difference between genders is fairly small or unreliable in some areas of spatial ability such as mental folding (Voyer, Voyer, & Bryden, 1995) and spatial perception (Linn & Petersen, 1985), this discrepancy is considerable in other areas such as mental rotation

(Hyde, 2014; Voyer, Voyer, & Bryden, 1995) and spatial visualization (Battista, 1990). The difference in spatial reasoning ability between genders appears during childhood and compounds throughout adolescence and adulthood (Geiser, Lehmann, & Eid, 2008; Lauer, Yhang, & Lourenco, 2019; Linn & Petersen, 1985).

Spatial Reasoning Assessments

Components of spatial reasoning are often assessed independently using specialized instruments. Mental rotation ability is commonly assessed using a variation of the Mental Rotation Test (Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). Measures of spatial orientation include the Perspective Taking Test (Hegarty, Kozhevnikov, & Waller, 2008), the Picture Test (Hegarty & Waller, 2004), and the Card Rotation and Cube Comparison Tests (Ekstrom & Harman, 1976). Spatial visualization assessments include the Revised Minnesota Paper Form Board Test (Quasha & Likert, 1937), the Mental Paper Folding Test (Shepard & Feng, 1972), and the Surface Development Test (Olkun, 2003). Spatial reasoning can also be measured using a single instrument with separate subscales for the different components (Ramful, Lowrie, & Logan, 2016). Traditional spatial reasoning instruments are typically administered as paper-and-pencil tests with clearly defined answers. This rigid structure limits the reusability of the instruments and does not allow students to generate their own solutions. These standardized instrument also suffer from a variety of problems commonly associated with this style of testing, such as an overemphasis on scores (Haladyna, Haas, & Allison, 1998; Sacks, 2000), artificially decreased scores due to test anxiety (Cassady & Johnson, 2002; Crocker, Schmitt, & Tang, 1988), and lower motivation and engagement as compared to other forms of assessments (Chiang & Lee, 2016; Papastergiou, 2009; Prensky, 2003).

Current Study

Shadowspect, an online geometry puzzle game, has players attempt to recreate silhouettes from different perspectives of a platform by manipulating and relating 3D primitive geometric shapes. Shadowspect was designed in part to evaluate players' spatial reasoning, with the long-term goal of serving as an ongoing assessment tool for educators. Shadowspect was modelled off of the math core-curriculum and deliberately designed with the needs of middle- and high-school instructors in mind, making it easier to integrate into lesson plans (Kim & Ruipérez-Valiente, 2020). Unlike the majority of existing spatial reasoning metrics, Shadowspect allows players to develop and test their own solutions to each puzzle. Shadowspect is also structured in such a way that a single puzzle may have multiple equally valid solutions, increasing the game's replay value.

The purpose of this study was to assess Shadowspect's potential as a spatial reasoning metric. Specifically, my research questions are:

RQ 1a: Which, if any, of Shadowspect's metrics are associated with spatial reasoning ability in general?

RQ 1b: Which, if any, of Shadowspect's metrics are associated with spatial reasoning components (mental rotation, spatial orientation, spatial visualization)?

Due to the potential differences between demographics, the study also aims to investigate:

RQ 2a: Does performance on Shadowspect vary among different age groups?

RQ 2b: If so, are these discrepancies also observed on the Spatial Reasoning Instrument?

RQ 3a: Does performance on Shadowspect vary among different gender groups?

RQ 3b: If so, are these discrepancies also observed on the Spatial Reasoning Instrument?

PROCEDURE

Participants

Fifty-four adults (19 men, 35 women) were recruited for this study using online recruitment flyers circulated via mailing lists, social media, and word of mouth. Two participants (1 man, 1 woman) were unable to complete the study due to technical difficulties and were not included in data analysis. The majority of participants (63.46%) were under the age of 36 years old, with 22 participants (42.31%) in the 18-25 age group and 11 participants (21.15%) in the 26-35 age group (see Table 1). The 36-45, 46-55, and 56+ age groups had 5 (9.62%), 9 (17.31%), and 5 (9.62%) participants respectively (see Table 1). Participants completed consent forms and background information surveys prior to their sessions via Qualtrics.

Table 1
Participant Demographics

	Gender		
Age	**M**	**F**	**Total**
18-25	7	15	22
26-35	4	7	11
36-45	2	3	5
46-55	2	7	9
56+	3	2	5

Note. This table does not include participants removed due to technical issues.

Procedure

This study was divided into two sections, (1) completing the Spatial Reasoning Instrument and (2) playing Shadowspect. The two sections were counterbalanced, with half of the participants completing the Spatial Reasoning Instrument portion first and half completing

the Shadowspect portion first. Each section took approximately thirty minutes, for a total session length of approximately one hour. Participants were given the option to complete both sections in one continuous session, with or without breaks, or in two separate sessions. Five participants chose to complete the sections in two separate sessions (three completed Spatial Reasoning Instrument portion in first session, two completed Shadowspect portion in first session).

Due to health and safety concerns associated with the COVID-19 pandemic, all sessions were proctored and conducted remotely using Zoom. At the beginning of sessions, participants were asked to share their screens. During a portion of the Shadowspect section, participants were asked permission to have their screen and audio recorded. Participants who had their screen and audio recorded were given the option of turning off their camera during the recording period.

Spatial Reasoning Instrument

Participants completed a digital version of Ramful, Lowrie, and Logan's (2016) paper-and-pencil Spatial Reasoning Instrument via Qualtrics. The questions in Ramful, Lowrie, and Logan's (2016) Spatial Reasoning Instrument were designed to recreate situations that middle school students would be familiar with and may encounter, especially in Science, Technology, Engineering, and Mathematics (STEM) classes, Shadowspect's target environment. An example of each question type can be found in Supplementary Figure 1**.**

The Spatial Reasoning Instrument contained subscales for the three spatial reasoning components of interest: mental rotation, spatial orientation, and spatial visualization. Ten questions were included for each construct for a total of 30 questions, each presented in a multiple choice format. All questions were equally weighted and worth one point for a maximum subscale score of 10 points and a maximum total score of 30 points. The amount of time

participants were given to complete the metric was decreased from 45 minutes to 30 minutes to account for the difference in age between the Spatial Reasoning Instrument's intended demographic and this study's adult participants to prevent the ceiling effect. All participants were able to complete the Spatial Reasoning instrument in the allotted thirty minutes.

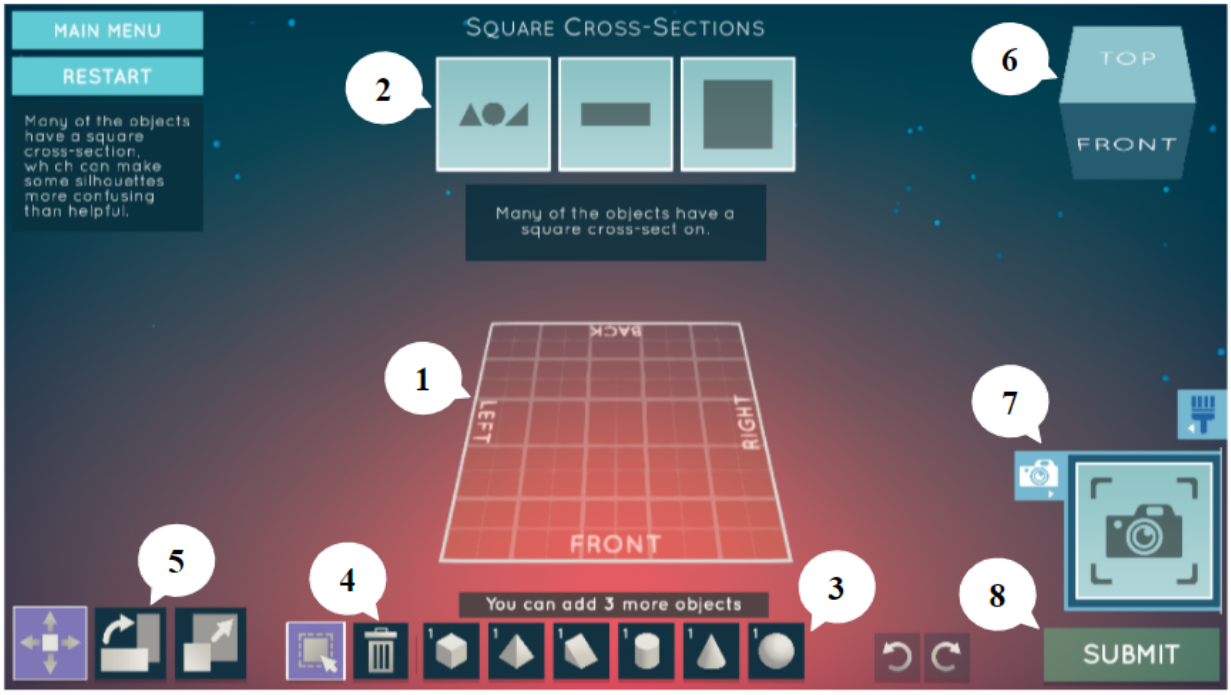
Shadowspect

Tutorial

During the Shadowspect portion of the study, participants completed a tutorial made up of three guided basic levels before playing five experimental levels unassisted. The three basic levels chosen for the tutorial were Separated Boxes, Rotate a Pyramid, and Stretch a Ramp. These tutorial levels were used to teach participants about Shadowspect's tools and mechanics. The objective of each level was to create a single arrangement using the shapes available in the Shape Menu in such a way that from different perspectives it matched a series of target images referred to as Shape Silhouettes (see Figure 1).

Figure 1

Shadowspect Layout



Note. (1) Platform, (2) Shape Silhouettes, (3) Shape Menu, (4) Delete Tool, (5) (left to right) Move, Rotate, and Stretch Buttons, (6) Perspective Cube, (7) Snapshot Tools, (8) Submit button.

In the first tutorial level, Separated Boxes, participants were taught how to add, move, and delete shapes; how to use the perspective cube to change their viewing angle; how to use the snapshot tool; how snapshots were resized; and how to identify and resolve collisions. A collision in Shadowspect was defined as one shape partially or completely intersecting another. Collisions resulted in both shapes being outlined in purple and submissions being disabled. In Rotate a Pyramid, participants were taught how to use the rotate button to rotate shapes and how to use the move and rotate buttons to switch between the moving and rotating arrows. In Stretch a Ramp, participants were taught how to use the stretch button to stretch and shrink a shape and how to use the move and stretch buttons to switch between moving and rotating arrows.

The tutorial levels were not timed, but participants were prompted after 30s-60s without visible progress. Verbal prompts included:

- “Do you have any questions about the tools?”
- “What do you think you could try next?”
- “Would you like a hint?”

In order to successfully complete the tutorial and proceed to the experimental levels, participants had to demonstrate understanding on at least one tutorial level. Understanding was defined as either (1) completing the level without assistance or (2) receiving assistance on the initial arrangement setup but being able to point out the platform perspective that matched each Shape Silhouette without assistance. Participants were asked a series of questions at the end of each tutorial level they received help on to assess understanding, including:

- “Can you point out which platform perspectives the different Shape Silhouette images were taken from?”
- “Why in the first Shape Silhouette does the arrangement look like [description], but in the second Shape Silhouette the arrangement looks like [description]?” Repeated for the third Shape Silhouette when applicable.
- “If you were to move [shape] [direction], why would this arrangement (no longer) be a valid solution?”

Participants were split into two completion groups based on whether they were able to successfully complete (COMP) or Did Not Complete (DNC) the tutorial levels. Participants in the DNC group did not attempt any of the experimental levels, and the Shadowspect portion of their session was ended. Participants who struggled with the tutorial but were able to demonstrate understanding were given more time to practice with Shadowspect’s tools.

Experimental Levels

Five Shadowspect levels were chosen as experimental levels: Scaling Round Objects (Basic), Square Cross-Sections (Intermediate), 45-Degree Rotations (Intermediate), More Than Meets Your Eye (Advanced), and Few Clues (Advanced). Across the five levels, all of the shapes available in the Shape Menu were used at least once (see Figure 1). Participants were told in advance that the Basic level would restrict which shapes and tools could be used and that Intermediate levels would give a maximum number of shapes but would not necessarily have specific shape restrictions (see Figure 2). Participants were also informed that Advanced levels would have no shape type or number restrictions and that there may be more than one possible correct arrangement.

Figure 2
Button and Shape Limitations for Each Level

Note. Any limitations on the buttons, type of shapes, or number of shapes that participants could use on a level were displayed at the bottom of each level. Restricted buttons were not visible to players on that level. Available shapes were given a blue background in the Shape Menu. The number of shapes that could be added appeared above the image of the shape in the Shape Menu.

Participants were given five minutes (300s) to complete each experimental level and were given warnings at the halfway mark and with one minute remaining. If the participant did not complete the level in the allotted five minutes but had an idea they wished to continue working on, they

were allowed up to an additional four minutes (240s) for a total of 540s. Participants were permitted to submit multiple times without penalty.

Analysis Plan

Independent-sample t-tests were used to compare differences in Spatial Reasoning Instrument scores, the number of levels solved, and the average solve time between genders. A t-test was deemed sufficient because all participants identified as either male or female (no participants selected “non-binary”, “other”, or “Prefer Not to Answer”). An independent-sample t-test was also used to compare the difference in overall Spatial Reasoning Instrument scores between completion groups.

Chi-square tests of independence were conducted to determine if there was an association between age group and Shadowspect tutorial completion, as well as if there was an association between gender and Shadowspect tutorial completion.

As a first step toward establishing Convergent Validity with a validated measure of spatial reasoning, correlations were calculated for the following variables: number of levels solved, solve time for each of the five experimental levels, average solve time, Spatial Reasoning Instrument overall score, mental rotation subscore, spatial orientation subscore, and spatial visualization subscore.

RESULTS

Fifty-two participants took the Spatial Reasoning Instrument and attempted all three levels of the Shadowspect tutorial. All 52 participants successfully completed the Spatial Reasoning Instrument, as measured by completing all 30 questions in the given 30 minutes. 44 participants (17 men, 27 women) successfully completed the tutorial (COMP) and attempted the experimental levels (see Table 2). Participants were grouped into either the DNC completion group if they were unable to complete Shadowspect's tutorial or the COMP completion group if they were successful.

Table 2

Participant Completion Groups Split by Gender

	Completion		
Gender	**DNC**	**COMP**	**Total**
M	1	17	18
F	7	27	34
Total	8	44	52

Note. Participants that successfully completed the tutorial were indicated by the abbreviation COMP. Participants that did not complete the tutorial were indicated by the abbreviation DNC.

Spatial Reasoning Instrument

The overall average Spatial Reasoning Instrument score across participants was 22.8 (*SD*=6.45) (see Table 3 and Figure 3). The maximum recorded score was 30 points, and the minimum was 8 points. The 18-25 age group had the highest average overall SRI score (*M*=26.1, *SD*=4.38), and the 46-55 age group had the lowest (*M*=14.3, *SD*=3.87). The difference in average overall scores between the highest scoring age group (18-25) and fourth highest scoring age

group (56+: *M*=20.6, *SD*=6.27) was less than the difference between the fourth highest scoring age group and the lowest scoring age group (46-55). The mental rotation (*M*=7.48, *SD*=2.87) and spatial visualization (*M*=6.44, *SD*=2.78) subscore distributions were similar to the overall score distribution. On the spatial orientation subscale, the majority of participants hit or approached the maximum subscore of 10 (*M*=8.90, *SD*=1.49). All participants answered at least one question correctly from each of the three subscales. Spatial Reasoning Instrument overall scores and subscores by age are shown in Supplementary Table 1.

Table 3

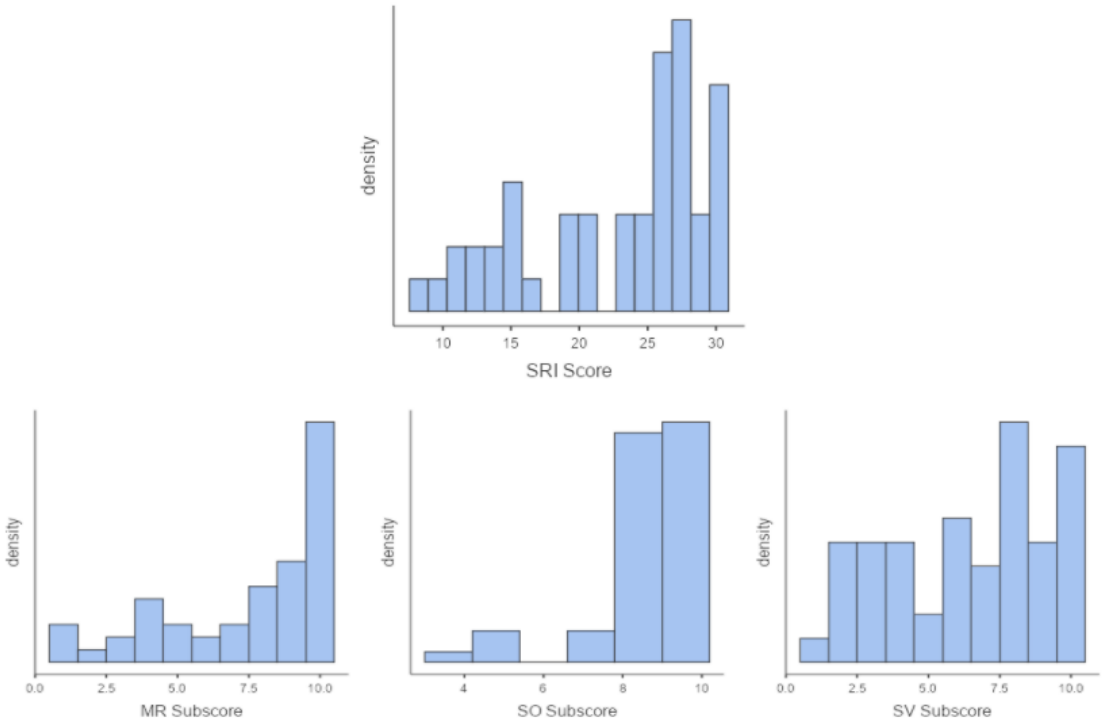
Spatial Reasoning Instrument (SRI) Scores and Subscores

	SRI Score	MR Subscore	SO Subscore	SV Subscore
N	52	52	52	52
Mean	22.8	7.48	8.90	6.44
Median	26.0	9.00	9.00	7.00
Standard deviation	6.45	2.87	1.49	2.78
Minimum	8	1	4	1
Maximum	30	10	10	10

Note. Mental Rotation, Spatial Orientation, and Spatial Visualization have been abbreviated to MR, SO, and SV respectively. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points. Possible MR, SO, and SV subscores ranged from 0-10 points each.

Figure 3

Spatial Reasoning Instrument Overall Score and Subscore Distributions



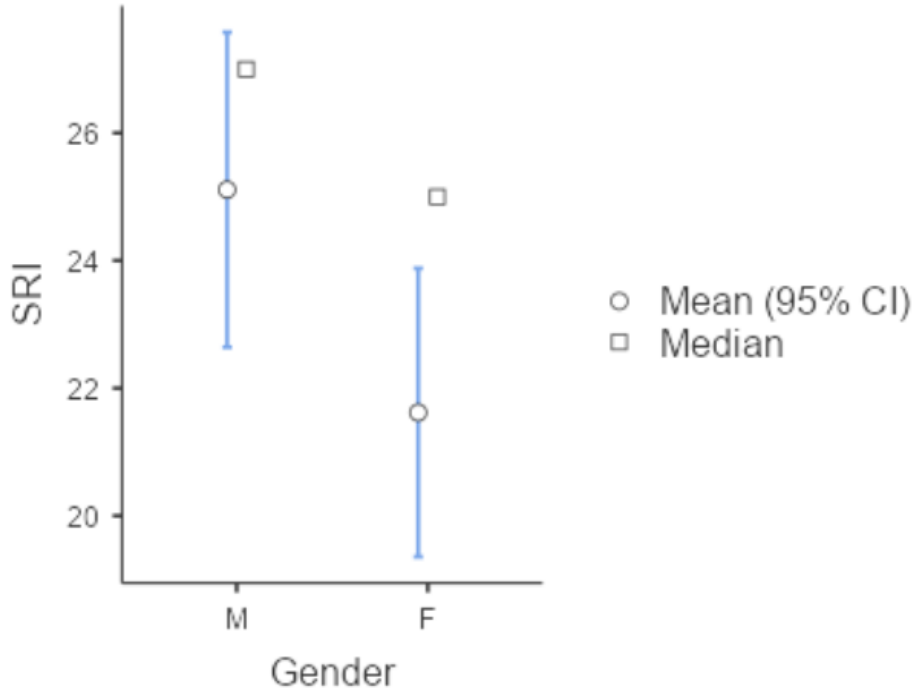
Note. Overall Spatial Reasoning Instrument distribution (*top*). Possible scores ranged from 0-30. Mental rotation subscale distribution (*bottom left*), spatial orientation subscale distribution (*bottom center*), spatial visualization subscale distribution (*bottom right*). Possible scores ranged from 0-10 for each subscale.

Gender

An independent-samples t-test was conducted to compare scores on the Spatial Reasoning Instrument between genders. There was no significant difference in Spatial Reasoning Instrument scores between genders; $t(50)=190$, $p=.063$ (see Figure 4).

Figure 4

Spatial Reasoning Instrument Score Split by Gender



Note. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points. Participant scores ranged from 8-30 points. Male and female participants scored on average 25.1 (*SD*=5.35) and 21.6 (*SD*=6.73) points respectively.

Shadowspect

All 44 participants in the COMP completion group attempted to solve all five experimental levels. An experimental level was considered solved if the participant assembled and submitted a valid arrangement in the allotted time without receiving any assistance. Participants were permitted to ask questions about Shadowspect's tools and mechanics such as how to switch between arrow types. Scaling Round Objects (Level 1) had the highest completion rate at 77.8% (N=35) followed by 45-Degree Rotations (Level 3) (57.8%, *N*=26), Square Cross-Sections (Level 2) (48.9%, *N*=22), More Than Meets Your Eye (Level 4) (46.7%, *N*=21), and Few Clues (Level 5) (15.6%, *N*=7) (see Table 4).

Table 4

Experimental Levels Attempted Solve Times

	Level 1	Level 2	Level 3	Level 4	Level 5
N	44	44	44	44	44
Missing	8	8	8	8	8
Solved (%)	35 (77.8%)	22 (48.9%)	26 (57.8%)	21 (46.7%)	7 (15.6%)
Mean (s)	226	377	321	404	546
Median (s)	148	440	234	600	600
Standard deviation (s)	206	231	244	214	130
Minimum (s)	39.0	59.0	48.0	80.0	127
Maximum (s)	600	600	600	600	600

Note. The 44 participants who successfully completed Shadowspect's tutorial (COMP) attempted the five experimental levels: Scaling Round Objects (Level 1), Square Cross-Section (Level 2), 45-Degree Rotations (Level 3), More Than Meets Your Eye (Level 4), and Few Clues (Level 5). For each level, COMP participants either (1) completed the level in the allotted 540s and were assigned that time for that level or (2) did NOT complete the level and were assigned a time of 600s for that level. DNC participants did not move onto the experimental levels, did not receive a time, and were therefore not included.

Table 5 shows the times of only the participants who successfully solved the levels either in the initial five-minute window or during the subsequent extension period. Among participants who solved the levels, Level 3 had the fastest completion time (T) (T=127s, SD=86s) followed by Level 1 (T=130s, SD=84s), Level 2 (T=254s, SD=69s), Level 4 (T=189s, SD=76s), and Level 5 (T=264s, SD=103s). No participants solved Level 1 or Level 2 after the initial five-minute window. Two participants solved Level 3 (N=26, 7.69%) after the initial five-minute window and averaged a solve time of 310 seconds. Two participants solved Level 4 (N=21, 9.52%) after the initial five-minute window and averaged a solve time of 375 seconds. Three participants solved Level 5 (N=7, 42.86%) after the initial five minutes and averaged a solve time of 408 seconds.

Table 5

Experimental Levels Solve Times

	Level 1	Level 2	Level 3	Level 4	Level 5
N	35	22	26	21	7
Missing	17	30	26	31	45
Mean (s)	130	154	127	189	264
Median (s)	106	145	85	179	235
Standard deviation (s)	84	69	86	76	103
Minimum (s)	39	59	48	80	127
Maximum (s)	298	280	310	375	430
# >300 (s)	0	0	2	2	3
AVG >300 (s)	0	0	10	75	108

Note. The 44 participants who successfully completed Shadowspect's tutorial (COMP) attempted the five experimental levels: Scaling Round Objects (Level 1), Square Cross-Section (Level 2), 45-Degree Rotations (Level 3), More Than Meets Your Eye (Level 4), and Few Clues (Level 5). Only COMP participants who completed each level in the allotted 540s were included in this table. Participants who were unsuccessful in solving a level in the allotted 540s were removed. The maximum solve time per level is 540s.

Age

Of the eight participants in the DNC group, one was in the 26-35 age group, six were in the 46-55 age group, and one was in the 56+ age group (see Table 6). For more information about the DNC completion group, see Supplementary Table 2. Six of the nine participants aged 46-55 (66.67%) did not complete the tutorial while all participants in the 18-25 (N=22) and 36-45 (N=5) age groups successfully completed the tutorial and progressed to the experimental levels. A chi-square test of independence was used to determine if there was an association between age group and Shadowspect tutorial completion. Age group was significantly associated with Shadowspect completion; $\chi^2(4) = 23.5,\ p < .001$.

Table 6

Participant Completion Groups Split by Age

Age	Completion		Total
	DNC	COMP	
18-25	0	22	22
26-35	1	10	11
36-45	0	5	5
46-55	6	3	9
56+	1	4	5
Total	8	44	52

Note. Participants that successfully completed the tutorial were indicated by the abbreviation COMP. Participants that did not complete the tutorial were indicated by the abbreviation DNC.

Average solve times (*T*) did not differ significantly among participants of different age groups, $F(4, 9.15)=2.63$, $p=0.104$. Average solve times for participants in the 18-25 (T=332s), 26-35 (T=370s), and 36-45 (T=399s) age groups were lower than those of participants in the 46-55 (T=525s) and 56+ (T=478s) groups (see Table 7**)**. The average number of levels differed significantly among participants of different age groups, $F(4, 14.2) = 9.86$, $p < .001$. Post-hoc analysis with Tukey adjustment reveals that the significance is driven by the difference between age groups 18-25 and 46-55 (M difference = 2.53, $p <.001$) as well as the difference between age groups 26-35 and 46-55 (M difference = 2.12, $p = .024$). The average number of levels solved was also higher for the 18-25 (M=2.86, SD=1.61), 26-35 (M=2.45, SD=1.86), and 36-45 (M=2.40, SD=1.34) age groups than the 46-55 (M=0.333, SD=0.707) and 56+ (M=1.20, SD=1.30) groups (see Table 7).

Table 7

Average (AVG) Solve Time and Number (#) of Levels Solved Split by Age

	Age	AVG Solve Time	# Levels Solved
N	18-25	22	22
	26-35	10	11
	36-45	5	5
	46-55	3	9
	56+	4	5
Missing	18-25	0	0
	26-35	1	1
	36-45	0	0
	46-55	6	6
	56+	1	1
Mean	18-25	332	2.86
	26-35	370	2.45
	36-45	399	2.40
	46-55	525	0.333
	56+	478	1.20
Median	18-25	284	3.00
	26-35	382	2
	36-45	368	3
	46-55	540	0
	56+	497	1
Standard deviation	18-25	155	1.61
	26-35	162	1.86
	36-45	133	1.34
	46-55	83.8	0.707
	56+	121	1.30
Minimum	18-25	117	0
	26-35	184	0
	36-45	228	1
	46-55	434	0
	56+	316	0
Maximum	18-25	600	5
	26-35	600	5
	36-45	542	4
	46-55	600	2
	56+	600	3

Note. Participants in the DNC group were not included. Solve time was measured in seconds (s).

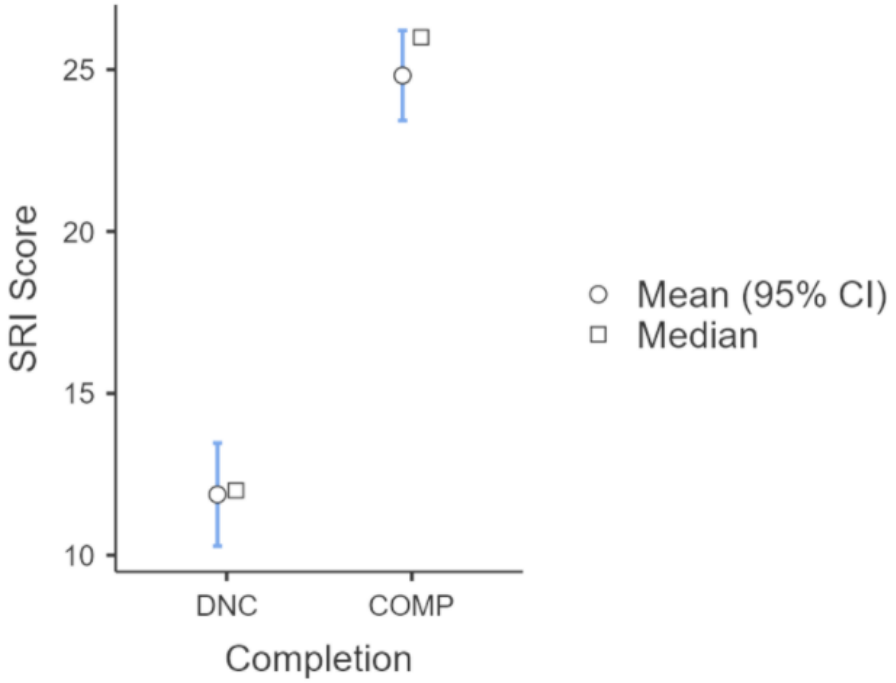
Gender

Of the eight participants in the DNC completion group, one was male and seven were female (see Table 2). A chi-square test of independence was used to determine if there was an association between gender and Shadowspect tutorial completion. Gender was not found to be significantly associated with Shadowspect tutorial completion; $\chi^2(1) = 2.04, p = .153$.

Spatial Reasoning Instrument & Shadowspect

An independent-samples t-test was conducted to compare overall scores on the Spatial Reasoning Instrument between the completion groups (DNC and COMP). There was a significant difference in Spatial Reasoning Instrument overall scores between completion groups; $t(50)=-7.57$, $p<.001$, Cohen's $d=-2.91$ (see Figure 5). The average overall Spatial Reasoning Instrument score for participants in the DNC group ($M=11.9$, $SD=2.30$) was less than half the average score of participants in the COMP group ($M=24.8$, $SD=4.71$) (see Table 8). The mental rotation subscale had the largest difference in subscores between the DNC ($M=2.63$, $SD=1.60$) and COMP ($M=8.36$, $SD=2.05$) groups. The smallest difference in subscores was on the spatial orientation subscale (DNC: $M=6.38$, $SD=1.85$; COMP: $M=9.36$, $SD=0.810$). Spatial Reasoning Instrument overall scores split by age and completion group is shown in Supplementary Table 3.

Figure 5
Spatial Reasoning Instrument Score Split by Completion



Note. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points. Participant scores ranged from 8-30 points. Participants in the DNC group scored significantly lower on average (M= 11.9, SD=2.30) than participants in the COMP group (M=24.8, SD=4.71).

Table 8

Spatial Reasoning Instrument Overall Score and Subscores Split by Completion

	Completion	SRI Score	MR Subscore	SO Subscore	SV Subscore
N	DNC	8	8	8	8
	COMP	44	44	44	44
Mean	DNC	11.9	2.63	6.38	2.88
	COMP	24.8	8.36	9.36	7.09
Median	DNC	12.0	2.50	6.00	3.00
	COMP	26.0	9.00	10.0	8.00
Standard deviation	DNC	2.30	1.60	1.85	1.25
	COMP	4.71	2.05	0.810	2.48
Minimum	DNC	8	1	4	1
	COMP	14	3	7	2
Maximum	DNC	15	5	9	4
	COMP	30	10	10	10

Note. Mental Rotation, Spatial Orientation, and Spatial Visualization have been abbreviated to MR, SO, and SV respectively. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points. Possible MR, SO, and SV subscores ranged from 0-10 points each.

Gender

In the COMP completion group, the average overall Spatial Reasoning Instrument scores for men (*N*=17) and women (*N*=27) were 25.8 (*SD*=4.54) and 24.2 (*SD*=4.78) respectively. (see Table 9). In the DNC group, the average score for men (*N*=1) was 13.0 (*SD*=NaN), and the average score for women (*N*=7) was 11.7 (*SD*=4.54).

Table 9

Spatial Reasoning Instrument (SRI) Overall Score Split by Gender and Completion

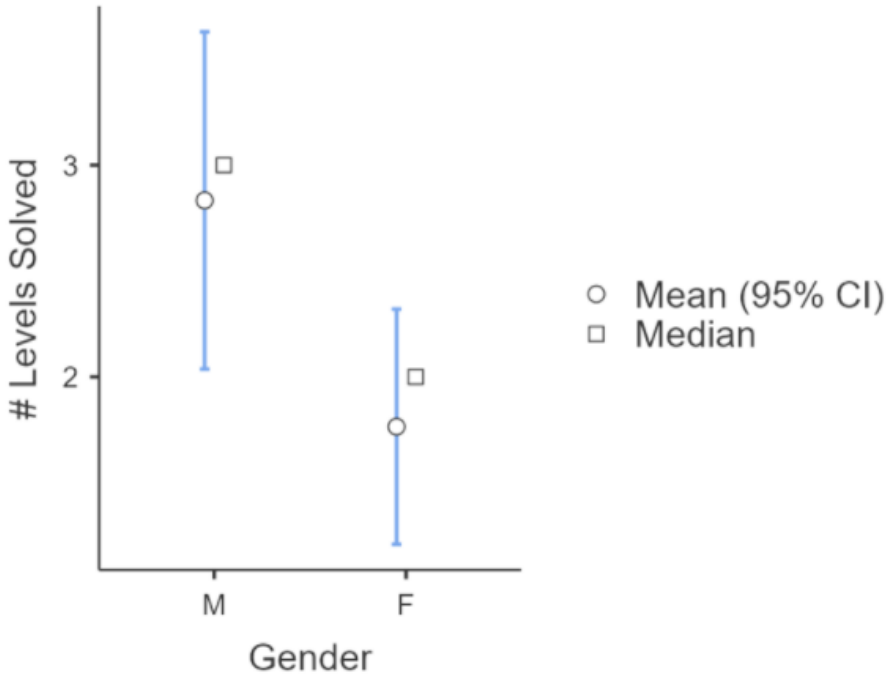
	Gender	Completion	SRI Score
N	M	DNC	1
		COMP	17
	F	DNC	7
		COMP	27
Mean	M	DNC	13.0
		COMP	25.8
	F	DNC	11.7
		COMP	24.2
Median	M	DNC	13
		COMP	28
	F	DNC	11
		COMP	26
Standard deviation	M	DNC	NaN
		COMP	4.54
	F	DNC	2.43
		COMP	4.78
Minimum	M	DNC	13
		COMP	15
	F	DNC	8
		COMP	14
Maximum	M	DNC	13
		COMP	30
	F	DNC	15
		COMP	30

Note. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points.

Independent-samples t-tests were conducted to compare the number of levels solved between genders and average level completion times between genders. There was a significant difference in the number of levels solved between the two genders; $t(50)=2.19$, $p=.034$, Cohen's $d=0.637$ (see Figure 6). Men solved on average one more level ($M=2.83$, $SD=1.72$) than women ($M=1.76$, $SD=1.65$). The levels with the greatest rate of solving difference between genders were Level 2 (M=71%, W=37%) and Level 4 (M=71%, W=33%).

Figure 6

Number of Levels Solved Split by Gender

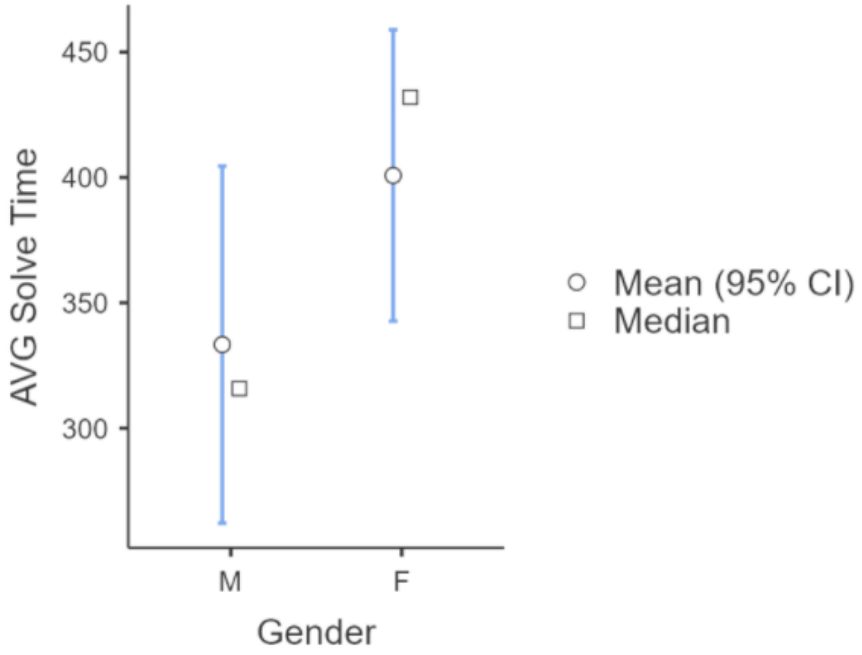


Note. Participants attempted to solve five levels. On average, men solved 2.83 ($SD=1.72$) levels, while women solved 1.76 ($SD=1.65$)

There was not a significant difference in average solve time between the two genders; $t(42)=-1.43$, $p=.160$ (see Figure 7).

Figure 7

Average Solve Time Split by Gender



Note. Solve time was measured in seconds (s). Participants had up to 540s to solve each level. Participants who were unable to solve a level were assigned a completion time of 600s for that level. Men averaged a solve time of 333s (*SD*=150), and women averaged a solve time of 401s (*SD*=154).

Correlations

As expected, average solve time was strongly and negatively correlated with the number of levels solved (r=-0.986, p<.001) (see Figure 8). The average time spent solving levels was strongly and negatively correlated with overall Spatial Reasoning Instrument scores (r=-0.579, p<.001) and the mental rotation (r=-0.530, p<.001) and spatial visualization (r=-0.520, p<.001) subscores. Average solve time was also negatively correlated to spatial orientation subscores (r=-0.437, p=.003) but more weakly than with the other subscales. The number of levels solved strongly and positively correlated with both overall scores on the Spatial Reasoning Instrument

(r=0.705, p<.001) and with subscores on the mental rotation (r=0.674, p<.001), spatial orientation (r=0.560, p<.001), and spatial visualization (r=0.640, p<.001) subsections.

Table 10
Correlation Matrix

		#_Levels_Solved	Time_Solved_LVL1	Time_Solved_LVL2	Time_Solved_LVL3	Time_Solved_LVL4	Time_Solved_LVL5	AVG_Solve_Time	SRI Score	MR Subscore	SO Subscore	SV Subscore
#_Levels_Solved	Pearson's r	—										
	p-value	—										
	N	—										
Time_Solved_LVL1	Pearson's r	-0.806***	—									
	p-value	< .001	—									
	N	44	—									
Time_Solved_LVL2	Pearson's r	-0.774***	0.497***	—								
	p-value	< .001	< .001	—								
	N	44	44	—								
Time_Solved_LVL3	Pearson's r	-0.704***	0.622***	0.275	—							
	p-value	< .001	< .001	0.07	—							
	N	44	44	44	—							
Time_Solved_LVL4	Pearson's r	-0.805***	0.584***	0.603***	0.504***	—						
	p-value	< .001	< .001	< .001	< .001	—						
	N	44	44	44	44	—						
Time_Solved_LVL5	Pearson's r	-0.555***	0.301*	0.304*	0.235	0.418**	—					
	p-value	< .001	0.047	0.045	0.125	0.005	—					
	N	44	44	44	44	44	—					
AVG_Solve_Time	Pearson's r	-0.986***	0.825***	0.737***	0.744***	0.844***	0.53***	—				
	p-value	< .001	< .001	< .001	< .001	< .001	< .001	—				
	N	44	44	44	44	44	44	—				
SRI Score	Pearson's r	0.705***	-0.353*	-0.358*	-0.521***	-0.555***	-0.351*	-0.579***	—			
	p-value	< .001	0.019	0.017	< .001	< .001	0.02	< .001	—			
	N	52	44	44	44	44	44	44	—			
MR Subscore	Pearson's r	0.674***	-0.312*	-0.316*	-0.512***	-0.499***	-0.307*	-0.53***	0.942***	—		
	p-value	< .001	0.039	0.037	< .001	< .001	0.043	< .001	< .001	—		
	N	52	44	44	44	44	44	44	52	—		
SO Subscore	Pearson's r	0.56***	-0.247	-0.293	-0.506***	-0.37*	-0.122	-0.437**	0.792***	0.673***	—	
	p-value	< .001	0.105	0.053	< .001	0.013	0.431	0.003	< .001	< .001	—	
	N	52	44	44	44	44	44	44	52	52	—	
SV Subscore	Pearson's r	0.64***	-0.333*	-0.323*	-0.401**	-0.522***	-0.373*	-0.52***	0.924***	0.792***	0.608***	—
	p-value	< .001	0.027	0.033	0.007	< .001	0.013	< .001	< .001	< .001	< .001	—
	N	52	44	44	44	44	44	44	52	52	52	—

Note. * p < .05, ** p < .01, *** p < .001

Spatial Reasoning Instrument scores were strongly and negatively correlated with solve times for Level 3 (r=-0.521, p<.001) and Level 4 (r=-0.555, p<.001). There was a weaker negative correlation between Spatial Reasoning Instrument scores and solve times for Level 1 (r=-0.353, p=.019), Level 2 (r=-0.358, p=.017), and Level 5 (r=-0.351, p=.020). Subscores on the mental rotation subscale were negatively correlated with solve times for all five levels, but especially Level 3 (r=-0.512, p<.001) and Level 4 (r=-.499, p<.001). Spatial orientation subscores were negatively correlated with solve times for Level 3 (r=-.506, p<.001) and Level 4 (r=-.370, p=.013), but were not correlated with solve times for Level 1 (p=.105) or Level 2 (p=.053). Spatial visualization subscores were negatively correlated with solve times for all five levels, but especially Level 3 (r=-0.401, p=.007) and Level 4 (r=-0.522, p<.001).

There was moderate to strong positive correlation between solve times amongst all levels except for Level 2 and Level 3 (p=.070) and Level 3 and Level 5 (p=.125). Solve times on Level 1 and Level 3 were strongly and positively correlated (r=0.622, p<.001) as were solve times on Level 2 and Level 4 (r=0.603, p<.001).

DISCUSSION

This study sought to assess the educational geometry game Shadowspect as a potential measure of spatial reasoning. Fifty-four participants were enrolled in the study. Fifty-two participants completed Ramful, Lowrie, and Logan's (2016) Spatial Reasoning Instrument. This Spatial Reasoning Instrument was chosen because it has three independent subscales for each construct of interest and because the metric was designed for a similar demographic as Shadowspect - middle- and high-school age students. While average scores on the Spatial Reasoning Instrument trended towards ceiling, especially on the spatial orientation subscale, the distributions were sufficient for the purposes of this study.

Performance on Spatial Reasoning Instrument and Shadowspect

Overall, participants performed well on the Spatial Reasoning Instrument, particularly on the spatial orientation subscale. On the spatial orientation subscale, only 13 participants scored lower than a 9 out of 10. The spatial orientation questions may have been generally easier on average, and the spatial orientation construct was the only subscale to contain questions with fewer than four answer choices (see Supplementary Figure 2). The Spatial Reasoning Instrument was designed for a younger demographic than the participants in this study, so changing the time constraints from 45 to 30 minutes may have been insufficient in increasing the difficulty on the spatial orientation questions. Scores on the mental rotation and spatial visualization subscales, as well as the Spatial Reasoning Instrument overall, were more equally distributed.

Forty-four participants successfully completed Shadowspect's tutorial and attempted five experimental levels: Scaling Round Objects (Level 1), Square Cross-Sections (Level 2),

45-Degree Rotations (Level 3), More Than Meets Your Eye (Level 4), and Few Clues (Level 5). These 44 participants formed the COMP completion group.

The Spatial Reasoning Instrument was able to capture and differentiate between the performances of participants at the lower end of the spatial reasoning ability spectrum. Shadowspect had a higher detection floor and was unable to measure spatial reasoning ability to the same degree of specificity as the Spatial Reasoning Instrument. Eight participants were unable to complete Shadowspect's tutorial and did not progress to the experimental levels. These eight participants formed the DNC completion group. The additional cognitive load associated with initially learning how to play Shadowspect may have artificially decreased participants' spatial reasoning scores below Shadowspect's detection threshold (Chandler & Sweller, 1991; Mayer, 2005; Sweller, 2011). Potential methods for improving Shadowspect's sensitivity are (1) increasing the number of tutorial levels, (2) increasing the time spent going through tutorial levels, (3) allocating time for open, unguided play, (4) incorporating the tutorial levels directly into the assessment process.

Shadowspect as a Potential Measure of Spatial Reasoning

Shadowspect successfully provided accurate assessments of spatial reasoning ability. The inability to complete Shadowspect's tutorial was strongly indicative of lower spatial reasoning ability. Conversely, a greater number of levels completed was a strong indicator of higher spatial reasoning ability, both overall and across subscores. Faster average completion times were also correlated with greater spatial reasoning ability. Among the subscales, Shadowspect performance was strongly related to greater mental rotation ability and, to a lesser extent, spatial visualization. While there was some correlation between Shadowspect performance and spatial orientation, the

relationship was noticeably weaker than that of the other subscales. This difference is likely due to the limited distribution of scores on the spatial orientation subscale, i.e., most participants scored the maximum or close to the maximum on this subscale.

Performance on individual levels, as measured by solve time, were also accurate indicators of spatial reasoning ability, most notably on the mental rotation subscale. Better performance on levels that heavily relied on rotation to solve (Levels 3 & 4) corresponded with higher mental rotation subscores. Performance on Level 3 was the best indicator of mental rotation ability. This could be attributed to Level 3 requiring the most rotation out of the five levels, with four shapes requiring at least one rotation each. Performance on Level 1, where rotation was not necessary and the rotate button was disabled, was very weakly correlated with mental rotation ability (see Figure 2).

Shadowspect performance was the least indicative of performance on the spatial orientation subscale. Spatial orientation subscores were only correlated with performance on Level 3 and, to a much lesser extent, Level 4. Spatial visualization subscores were positively and most strongly correlated with performances on Level 3, Level 4, and Level 5. These level's Shape Silhouettes included images of shapes overlapping (Levels 3 & 5) and stacking (Level 4), defined as "impossible" composite shapes. In order to solve the levels, participants needed to mentally deconstruct these impossible shapes into the subparts corresponding to shapes available in the Shape menu.

Levels with positively correlated solve times had certain features in common including level difficulty, lack or presence of shape choice, and shape cross-sections. Generally, level performance was positively correlated between levels of progressive difficulty (Levels 1 & 2,

Levels 3 & 4, Levels 4 & 5). Performances were also positively correlated on levels that required participants to choose which shapes they used (Levels 2, 4, & 5) (see Figure 2). Level 1 and Level 3 solve times were the most strongly correlated amongst the levels, likely due to the common lack of choice in shape number or type. Participants only needed to arrange and manipulate the shapes to solve these levels. Level 2 and Level 4 were strongly correlated, likely due to both levels involving small shape rotations and requiring heavy use of square cross-sections. Performance on Level 3 was not related to performance on either Level 2 or Level 5, both of which required correct shape selection. These three correlation factors may be helpful when choosing which levels to include in future studies.

Age and Gender Differences

Age was related to performance on both the Spatial Reasoning Instrument and Shadowspect. Younger participants scored higher on average on the Spatial Reasoning Instrument than older participants. Younger participants also performed significantly better on Shadowspect as measured by tutorial completion rate, average level solve times, and number of levels solved. Spatial reasoning ability and speed have both been observed to decline in later adulthood (Cerella, Poon, & Fozard, 1981; Lord & Marsh, 1975). This difference is exacerbated when attempting to perform multiple concurrent spatial transformations concurrently (Salthouse, 1987). Moreover, the average age of video game players tends to skew younger, with children and young adults making up a higher percentage of players and playing more video games than middle-aged and older adults (Clement, 2021; Pew Research Center, 2017).Younger participants' familiarity with games may have decreased their cognitive load, resulting in a better overall performance (Chandler & Sweller, 1991; Mayer, 2005; Sweller, 2011).

Although men have traditionally outperformed women in measures of spatial reasoning (Halpern 2013; Linn & Petersen, 1985; Maccoby & Jacklin, 1974; Voyer, Voyer, & Bryden, 1995), there was no significant difference in the Spatial Reasoning Instrument scores between genders in the current study. There were also no significant differences in Shadowspect tutorial completion or average solve time between genders. Nevertheless, male participants solved on average one more level than female participants, a difference worth investigating. One possible explanation on initial examination is that overall, the number of levels solved correlated positively with Spatial Reasoning Instrument scores for both genders independently, regardless of the crude number of levels solved. In other words, men that solved more levels had higher Spatial Reasoning Instrument scores than men that solved fewer levels. This trend was then preserved when men and women's performances were combined, supported by the gender parity in Spatial Reasoning Instrument scores. The overall gender parity in tutorial completion and average solve time indicate that Shadowspect could be a useful tool for players of all genders.

Future Directions

Although adult participants were sufficient for this preliminary study, Shadowspect was intended for use in middle- and high-school classrooms. A follow-up study with a more diverse pool of participants from the target demographic is necessary to determine Shadowspect's effectiveness as a valid measure in its intended use environment. This study focused on the macro-action data collected by Shadowspect, particularly completion rate and solve times. Future work could focus on more specific, micro-action data such as the use of particular tools, platform rotation, and player shape preference as potential measures of spatial reasoning ability.

REFERENCES

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for research in mathematics education*, *21*(1), 47-60.

Ben-Chaim, D., Lappan, G., & Houang, R. T. (1988). The effect of instruction on spatial visualization skills of middle school boys and girls. *American Educational Research Journal*, *25*(1), 51-71.

Burnett, S. A., & Lane, D. M. (1980). Effects of academic instruction on spatial visualization. *Intelligence*, *4*(3), 233-242.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary educational psychology*, *27*(2), 270-295.

Cerella, J., Poon, L. W., & Fozard, J. L. (1981). Mental rotation and age reconsidered. *Journal of Gerontology*, *36*(5), 620-624.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, *8*(4), 293-332.

Cheng, Y. L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, *15*(1), 2-11.

Cherney, I. D. (2008). Mom, let me play more computer games: They improve my mental rotation skills. *Sex Roles*, *59*(11-12), 776-786.

Chiang, C. L., & Lee, H. (2016). The effect of project-based learning on learning motivation and problem-solving ability of vocational high school students. *International Journal of Information and Education Technology*, *6*(9), 709-712.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. *Handbook of research on mathematics teaching and learning*, 420-464.

Clement, J. (2021, May 5). *U.S. average age of video gamers 2019*. Statista. https://www.statista.com/statistics/189582/age-of-us-video-game-players-since-2010/.

Corradini, A. (2011). A study on whether digital games can effect spatial reasoning skills. In *Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches* (pp. 1086-1110). IGI Global.

Crocker, L., Schmitt, A., & Tang, L. (1988). Test anxiety and standardized achievement test performance in the middle school years. *Measurement and Evaluation in Counseling and Development*, *20*(4), 149-157.

De Lisi, R., & Wolford, J. L. (2002). Improving children's mental rotation accuracy with

computer game playing. *The Journal of genetic psychology*, *163*(3), 272-282.

Dorval, M., & Pepin, M. (1986). Effect of playing a video game on a measure of spatial visualization. *Perceptual and motor skills*, *62*(1), 159-162.

Ehrlich, S. B., Levine, S. C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental psychology*, *42*(6), 1259.

Ekstrom, R. B., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service.

Eraso, M. (2007). Connecting visual and analytic reasoning to improve students' spatial visualization abilities: A constructivist approach.

Gagnon, D. (1985). Videogames and spatial skills: An exploratory study. *ECTJ*, *33*(4), 263-275.

Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, *36*(6), 556-563.

Gorska, R., & Sorby, S. (2008, June). Testing instruments for the assessment of 3 D spatial skills. In *2008 Annual Conference & Exposition* (pp. 13-1196).

Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education*, *74*(5), 262-273.

Halpern, D. F. (2013). *Sex differences in cognitive abilities*. Psychology press.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*(2), 175-191.

Hegarty, M., Kozhevnikov, M., & Waller, D. (2008). Perspective taking/spatial orientation test. *University of California, Santa Barbara. Consultado el*, *5*.

Hsi, S., Linn, M. C., & Bell, J. E. (1997). The role of spatial reasoning in engineering and the design of spatial instruction. *Journal of engineering education*, *86*(2), 151-158.

Hyde, J. S. (2014). Gender similarities and differences. *Annual review of psychology*, *65*, 373-398.

Kim, Y. J., & Ruipérez-Valiente, J. A. (2020, September). Data-Driven Game Design: The Case of Difficulty in Educational Games. In *European Conference on Technology Enhanced Learning* (pp. 449-454). Springer, Cham.

Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological bulletin*, *145*(6), 537.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, 1479-1498.

Lohman, D. F. (1979). Spatial ability: A review and reanalysis of the correlational literature.

Lord, S. A. G., & Marsh, G. R. (1975). Age differences in the speed of a spatial cognitive process. *Journal of Gerontology*, *30*(6), 674-678.

Lowrie, T., Logan, T., & Hegarty, M. (2019). The influence of spatial visualization training on students' spatial reasoning and mathematics performance. *Journal of Cognition and Development*, *20*(5), 729-751.

Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. *Personality and Individual Differences*, *49*(4), 344-351.

Maccoby, E. E., & Jacklin, C. N. (1978). *The psychology of sex differences* (Vol. 2). Stanford University Press.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, *41*, 31-48.

McClurg, P. A., & Chaillé, C. (1987). Computer games: Environments for developing spatial cognition?. *Journal of educational computing research*, *3*(1), 95-111.

Olkun, S. (2003). Making connections: Improving spatial abilities with engineering drawing activities. *International journal of mathematics teaching and learning*, *3*(1), 1-10.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & education*, *52*(1), 1-12.

Pew Research Center. (2017, September 7). *Younger Americans and men are among the most likely to play video games*. Pew Research Center. https://www.pewresearch.org/fact-tank/2017/09/11/younger-men-play-video-games-but-so-do-a-diverse-group-of-other-americans/ft_17-09-11_videogames_youngeramericans/.

Pittalis, M., & Christou, C. (2010). Types of reasoning in 3D geometry thinking and their relation with spatial ability. *Educational Studies in mathematics*, *75*(2), 191-212.

Prensky, M. (2003). Digital game-based learning. *Computers in Entertainment (CIE)*, *1*(1), 21-21.

Quasha, W. H., & Likert, R. (1937). The revised Minnesota paper form board test. *Journal of Educational Psychology*, *28*(3), 197

Ramful, A., Lowrie, T., & Logan, T. (in press). Measurement of spatial ability: Construction and

validation of the spatial reasoning instrument for middle school students. *Journal of Psychoeducational Assessment.*

Sacks, P. (2000). Standardized minds: The high price of America's testing culture and what we can do to change it. *National Association of Secondary School Principals. NASSP Bulletin*, *84*(616), 118.

Salthouse, T. A. (1987). Adult age differences in integrative spatial ability. *Psychology and Aging*, *2*(3), 254.

Septia, T., & Prahmana, R. C. I. (2018). Improving Students Spatial Reasoning with Course Lab. *Journal on Mathematics Education*, *9*(2), 327-336.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive psychology*, *3*(2), 228-243.

Stieff, M., & Uttal, D. (2015). How much can spatial training improve STEM achievement?. *Educational Psychology Review*, *27*(4), 607-615.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, *47*(2), 599-604.

Velez, M. C., Silver, D., & Tremaine, M. (2005, October). Understanding visualization through spatial ability differences. In *VIS 05. IEEE Visualization, 2005.* (pp. 511-518). IEEE.

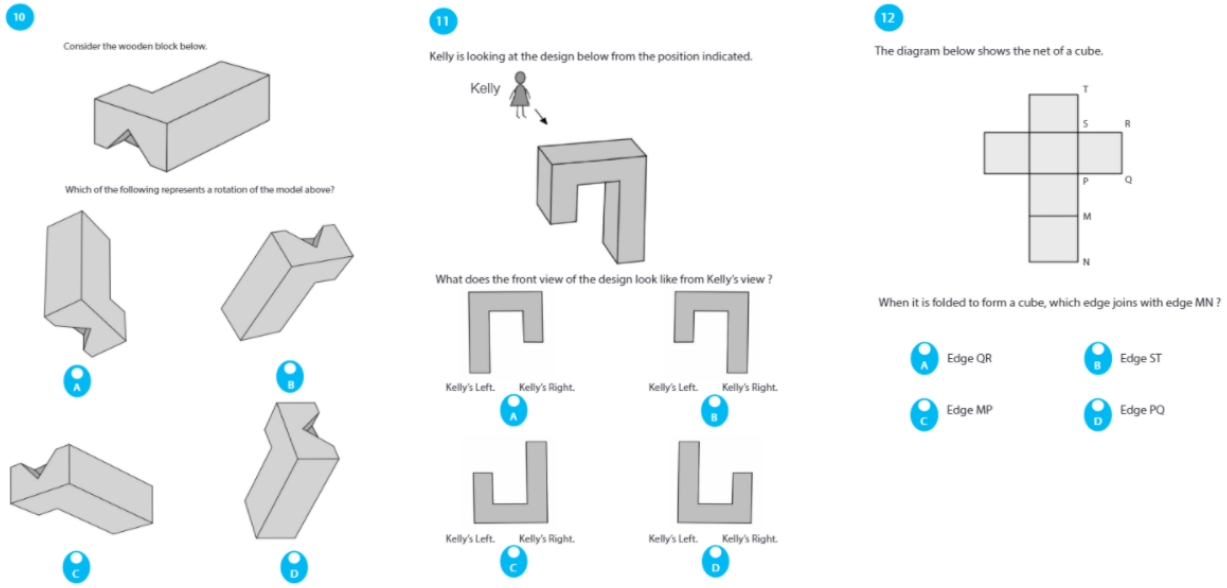
Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. (2017). *Links between spatial and mathematical skills across the preschool years*. Hoboken: Wiley.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological bulletin*, *117*(2), 250.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of educational psychology*, *101*(4), 817.

SUPPLEMENTARY MATERIAL

Supplementary Figure 1



Note. An example of a mental rotation (left), spatial orientation (center), and spatial visualization (right) question from the Spatial Reasoning Instrument.

Supplementary Table 1.
Spatial Reasoning Instrument (SRI) Overall Score and Subscores Split by Age

	Age	SRI Score	MR Subscore	SO Subscore	SV Subscore
N	18-25	22	22	22	22
	26-35	11	11	11	11
	36-45	5	5	5	5
	46-55	9	9	9	9
	56+	5	5	5	5
Mean	18-25	26.1	8.86	9.45	7.82
	26-35	23.1	7.55	8.82	6.73
	36-45	25.2	9.20	9.60	6.40
	46-55	14.3	3.67	7.33	3.33
	56+	20.6	6.40	8.80	5.40
Median	18-25	27.0	9.50	10.0	8.00
	26-35	25	9	9	6
	36-45	26	10	10	7
	46-55	14	3	8	4
	56+	19	5	9	6
Standard deviation	18-25	4.38	1.67	0.858	2.42
	26-35	6.79	2.94	1.60	2.69
	36-45	3.11	1.10	0.548	2.07
	46-55	3.87	2.35	2.06	1.32
	56+	6.27	2.88	1.10	2.88
Minimum	18-25	15	4	7	2
	26-35	8	1	5	2
	36-45	21	8	9	3
	46-55	10	1	4	1
	56+	13	4	7	2
Maximum	18-25	30	10	10	10
	26-35	30	10	10	10
	36-45	28	10	10	8
	46-55	21	8	9	5
	56+	28	10	10	9

Note. Mental Rotation, Spatial Orientation, and Spatial Visualization have been abbreviated to MR, SO, and SV respectively. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points. Possible MR, SO, and SV subscores ranged from 0-10 points each. The 18-25 age group had the highest average overall SRI score (M=26.1, SD=4.38) followed by 36-45 (M=25.1, SD=3.11), 26-35 (M=23.1, SD=6.79), 56+ (M=20.6, SD=6.27), and 46-55 (M=14.3, SD=3.87).

Supplementary Table 2

DNC Participants Demographic and Spatial Reasoning Instrument (SRI) Information

Participant ID#	Age	Gender	SRI Score
001	56+	M	13
007	46-55	F	13
015	46-55	F	11
016	46-55	F	11
017	46-55	F	14
032	46-55	F	15
037	26-35	F	8
042	46-55	F	10

Note. Eight participants did not complete Shadowspect's tutorial. Average SRI score for those participants was 11.9 (*SD*=2.30) points.

Supplementary Table 3
Spatial Reasoning Instrument (SRI) Scores Split by Age and Completion

	Completion	Age	SRI Score
N	DNC	18-25	0
		26-35	1
		36-45	0
		46-55	6
		56+	1
	COMP	18-25	22
		26-35	10
		36-45	5
		46-55	3
		56+	4
Mean	DNC	18-25	NaN
		26-35	8.00
		36-45	NaN
		46-55	12.3
		56+	13.0
	COMP	18-25	26.1
		26-35	24.6
		36-45	25.2
		46-55	18.3
		56+	22.5
Median	DNC	18-25	NaN
		26-35	8
		36-45	NaN
		46-55	12.0
		56+	13
	COMP	18-25	27.0
		26-35	25.5
		36-45	26
		46-55	20
		56+	22.5
Standard deviation	DNC	18-25	NaN
		26-35	NaN
		36-45	NaN
		46-55	1.97
		56+	NaN

Minimum	DNC	18-25	NaN
		26-35	8
		36-45	NaN
		46-55	10
		56+	13
	COMP	18-25	15
		26-35	15
		36-45	21
		46-55	14
		56+	17
Maximum	DNC	18-25	NaN
		26-35	8
		36-45	NaN
		46-55	15
		56+	13
	COMP	18-25	30
		26-35	30
		36-45	28
		46-55	21
		56+	28

Note. Possible scores on the Spatial Reasoning Instrument ranged from 0-30 points.

Supplementary Figure 2

2

Kate and William's seating positions are shown below.

In which position is the flower vase from Kate's view?

A To her right.

B To her left.

5

The diagram below shows a ballet dancer facing the audience.

The ballet dancer is extending one of her arms.
Which arm has she extended?

A Right

B Left

Note. Two questions on the spatial orientation subscale from the Spatial Reasoning Instrument had fewer than four answer choices.