Draft version December 11, 2019
Typeset using LaTeX default style in AASTeX63

The Effectiveness of the ThinkSpace Curriculum on Student Learning in Middle School Astronomy

Nicole Gordon[1] and Dr. Patricia Udomprasert[2]

[1] *Harvard University*
[2] *Harvard University*
60 Garden Street, MS 42
Cambridge, MA, 02318, USA

ABSTRACT

ThinkSpace is a middle school astronomy curriculum for teaching seasons with activities that require students to make use of spatial strategies. The materials for the curriculum involve hands on models, visualizations, and computer simulations. Students are asked to use these tools to visualize different aspects of the Earth-Sun system and relate them to the cause of seasons. The aim of this research is to determine how effective the ThinkSpace curriculum is compared to a traditional seasons curriculum that does not focus on spatial thinking. This paper uses data from one school district in which the curriculum was implemented. The ThinkSpace curriculum was originally administered when the students were in sixth grade. In the district, about one fourth of the sixth graders used the ThinkSpace curriculum to learn seasons while the remaining sixth grade students used the traditional seasons curriculum. The students who used ThinkSpace in sixth grade were given a pre-test and a post-test immediately before and after ThinkSpace instruction, respectively. Two years later, when the students were in eighth grade, the district moved their astronomy requirement from sixth grade to eighth grade, so the same students had to take astronomy again. This time, the district mandated all students use the ThinkSpace curriculum. A pre-test was administered immediately prior to eighth-grade instruction. By comparing the eighth-grade pre-test scores of the students who used ThinkSpace in sixth grade to the eighth-grade pre-test scores of the students who did not ThinkSpace in sixth grade, we can measure the effectiveness of the ThinkSpace curriculum on long-term retention. Using a t-test we find that out of eight questions, the eighth-grade pre-test scores are $0.76 \pm .13$ ($p < 0.0001$) higher for the students who used ThinkSpace in sixth grade compared to the students who did not use ThinkSpace in sixth grade. We also find an effect size of 0.54 ± 0.01 when comparing the mean score on the eighth-grade pre-test of the sixth-grade ThinkSpace students to the sixth-grade non-ThinkSpace students. From these results we conclude that in the long-term, the ThinkSpace curriculum is more effective than a traditional seasons curriculum.

1. INTRODUCTION

In a national study of 48 states, it was found that 46 states include the Earth's tilt/seasons in their core curriculum benchmarks for grades five through eight, with the average grade being grade six (Palen & Proctor (2006)). Although seasons is a very common topic in middle school science, understanding what causes seasons is challenging for students (Sneider et al. (2011)). There are multiple reasons for this, but there are two that significantly motivated the creation of the ThinkSpace seasons curriculum (Sneider et al. (2011)):

1. There are few materials used in teaching seasons that involve students observing the path of the Sun in the sky and how it changes throughout the year.
2. A high level of spatial reasoning skills is required to reconcile the Sun's changing path in the sky (the Earth-based perspective) with the rotation of the Earth on its tilted axis and its revolution around the Sun (the space-based perspective).

Corresponding author: Dr. Patricia Udomprasert
nicolegordon@college.harvard.edu, pudomprasert@cfa.harvard.edu

The second challenge presented above is especially important because it has been found that perspective taking is a spatial skill which appears to impact how well one understands astronomical phenomena (Liben & Downs (1993) and Plummer et al. (2016)). Perspective taking is being able to visualize events from different points of view. Based on this, it is recommended that science teachers amend their astronomy curriculum to include perspective-taking activities. Students can work to improve their spatial skills by using simulations and physical models that allow them to visualize a system from multiple viewpoints (Plummer et al. (2016)).

Another challenging aspect of learning seasons is that students often learn about seasons from textbook images which can be confusing and misleading. Astronomy is a field with three dimensional phenomena that are forced into two dimensional images. This means that the image must be shown from a specific point of view (usually from the side or from the top down). This leads to misleading images because the drawings are often complex, the symbols are ambiguous, and it is difficult to imagine what the image would look like in three dimensions (Galano et al. (2018)). One example of this is explaining the cause of seasons by using arrows to represent solar rays hitting Earth (See Figure 1). If the arrows extend all the way to the surface of the tilted Earth, it appears as if the Sun rays hitting the pole tilted away from the Sun travel significantly farther than the Sun rays hitting the pole tilted towards the Sun (Ojala (1992)). This could lead to the misconception that since the Earth is tilted, the Sun's rays travel different distances to the Northern and Southern hemispheres, causing summer in one and winter in the other (Galano et al. (2018)). Another

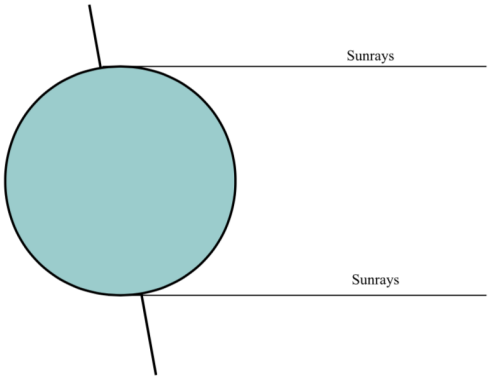


Figure 1. A representation of Sun rays hitting the Earth with reference to Earth's tilted axis. Credit to Galano et al. (2018) who adapted the image from Ojala (1992).

misleading textbook image is the orbit of the Earth around the Sun (See Figure 2). The Earth's orbit, which is only very slightly elliptical, is often depicted from the side, which exaggerates the eccentricity (Ojala (1992) and Galano et al. (2018)). Figure 1 and Figure 2 are not actual images from textbooks, but they are both common misrepresenta-

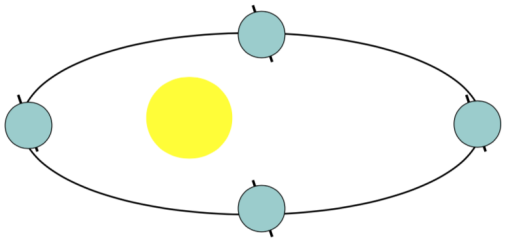


Figure 2. The orbit of the Earth around the Sun. Credit to Galano et al. (2018) who adapted the image from Ojala (1992).

tions that are often seen in textbooks. The images are from Galano et al. (2018), who adapted them from Ojala (1992).

The ThinkSpace curriculum was created to change the way middle school astronomy is taught by creating lessons based on the spatial aspect of astronomy. The goal of the ThinkSpace curriculum is for students to be able to

visualize astronomical phenomena and eliminate their misconceptions about seasons and related topics. It includes an eight-day seasons curriculum with demonstrations, hands-on student activities, videos, and worksheets that teachers can use. There is also a ThinkSpace curriculum on lunar phases, but the data analysed in this paper will only focus on the seasons curriculum. Many of the videos and models use WorldWide Telescope (WWT), an interactive computer program which creates visualizations of the Universe using images from telescopes all around the world. The ThinkSpace curriculum uses WWT to help students conceptualize spatial ideas such as the relative sizes of objects in the Solar Solar System, the shape of Earth's orbit, and how sunlight and the orientation of the Earth causes seasons.

The difference between the ThinkSpace curriculum and traditional seasons curricula is the emphasis on spatial thinking. The ThinkSpace curriculum includes activities with hands on and virtual models that require a student to use spatial reasoning. Students are asked to describe, from a person's perspective on Earth, how the Sun appears to move across the sky throughout the day and how this changes based on the time of year (Udomprasert et al. (2019)). They are then asked to describe, from a space-based perspective, the Earth's rotation on its tilted axis as it revolves around the Sun (Udomprasert et al. (2019)). The students must then reconcile these two perspectives and explain how the space-based description causes what is seen from Earth (Udomprasert et al. (2019)). These visualizations are done using WWT. The computer program allows the user to switch back-and-forth from an Earth-based perspective to a space-based perspective. From Earth, students can manipulate the time of day and year and track the Sun's actual path across the sky. Students explore how long and how high the Sun is in the sky at different points in the year (See Figure 3). They then plot the Sun's path at different times of the year on a Suntracker (See Figure 4), which was created by Dr. Philip Sadler.

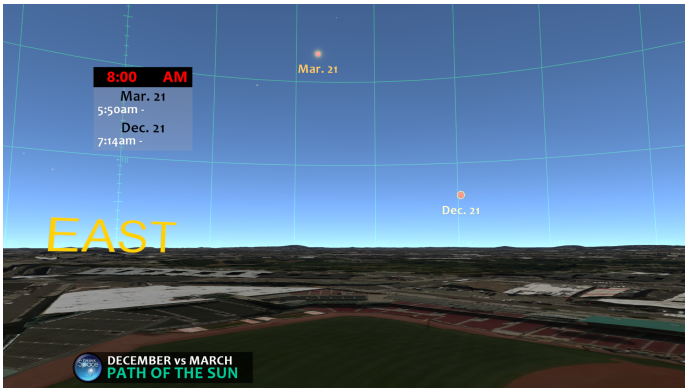


Figure 3. An image from WWT showing how students track the path of the Sun in the sky at different times in the year.

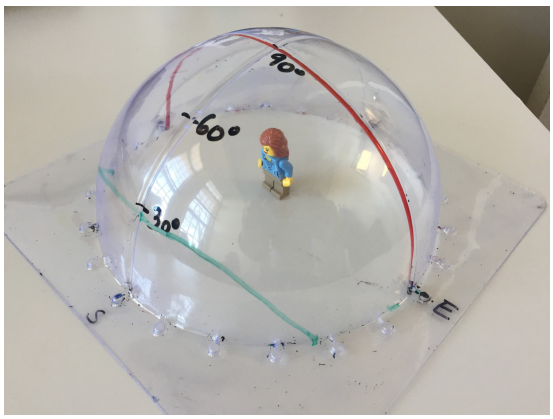


Figure 4. An image of a Suntracker, created by Dr. Philip Sadler, which students use to plot the path of the Sun in the sky at different times in the year.

From space, students can vary the time of year to discover how a point on the globe experiences different lengths of day and night depending on where the Earth is in its orbit (See Figure 5).

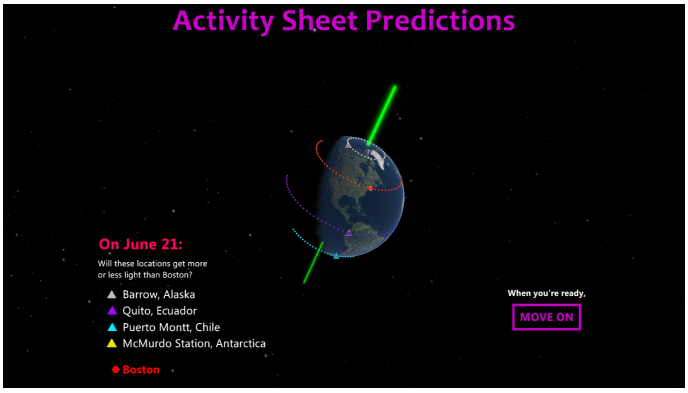


Figure 5. An image from an activity using WWT in which students observe the Earth from space; the students explore how the Earth's tilt affects the length of daylight specific locations experience.

Students can then compare the Earth- and space-based perspectives to understand how the Earth's tilt determines how much daylight a location experiences and how high the Sun appears to travel in the sky. Finally, WWT is used to explore the shape of Earth's orbit from different points of view (See Figure 6 and Figure 7). By being able to view the Earth's orbit from any angle, students are able to see that the Earth's orbit is not highly elliptical. This can help them dispel the misconception that the Earth's changing distance from the Sun is the cause of the seasons (Udomprasert et al. (2019)).

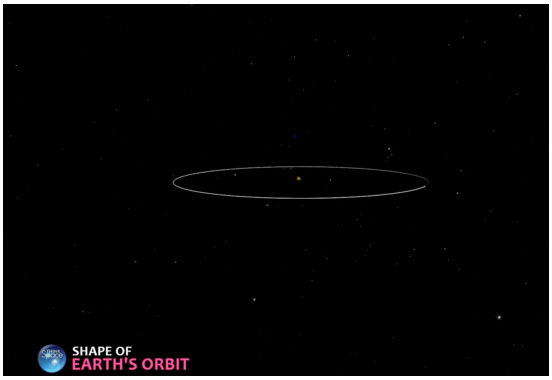


Figure 6. An image from WWT of the Earth's orbit around the Sun as viewed from the side. From this angle, the Earth's orbit appears to be highly elliptical.

Another motivation for creating a seasons curriculum based on spatial thinking comes from the result that a person's spatial abilities have a large impact on their achievement in STEM (Wai et al. (2009)). It was also found that people with strong spatial skills but more average mathematical and verbal skills have the potential to do well in a STEM field (Wai et al. (2009)). Spatial skills are also found to be malleable and are able to be improved (Uttal et al. (2013)). This implies increased spatial training causes improvement in a person's ability to learn material in STEM fields (Uttal et al. (2013)).

The popular misconceptions addressed in the ThinkSpace curriculum come from The Astronomy and Space Science Concept Inventory (ASSCI) published by Sadler et al. (2010). Sadler et al. (2010) researched common misconceptions and created distractor-driven tests that can be used to determine what misconceptions students hold (Sadler et al. (2010)). A distractor-driven test is a multiple-choice test where one or more of the answer choices are popular misconceptions. This forces the students to fully understand the concepts they are being tested on rather than just being able to recall keywords or choose the correct answer by a simple process of elimination. In their national

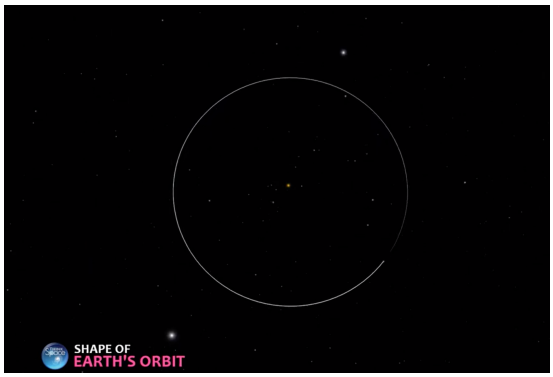


Figure 7. An image from WWT of the Earth's orbit around the Sun as viewed from the top. Students are able to manipulate the viewing angle of the orbit to see that Earth's orbit is only very slightly elliptical.

survey, they found that student gains in knowledge at the high school level (0.51 ± 0.05) are higher than at the middle school level (0.13 ± 0.03), but their research was inconclusive about why this happens (Sadler et al. (2010)). Due to their lack of concrete results on this topic, they acknowledge that there are opportunities to use the ASSCI tests as pre-tests and post-tests to determine what conditions lead to higher gains in student knowledge (Sadler et al. (2010)).

Based on this literature, creating a seasons curriculum that focuses on the use of spatial abilities could be beneficial for both student learning in the classroom and their future experiences in STEM. It is important to test the impact of the ThinkSpace curriculum by comparing results from students who use the curriculum to those who do not. The ASSCI instruments can be used as a pre-test before instruction and a post-test after instruction to determine how much more students learned from the spatially-based curriculum and what misconceptions were eliminated. The motivating research question of this paper is:

- Is the middle school ThinkSpace seasons curriculum more effective on long-term student learning than a traditional seasons curriculum?

2. METHODS

2.1. *Data Collection*

The data used in this paper was collected from two schools in a public school district in a suburban area in Massachusetts. These two schools were chosen for the study because they were particularly strong partner schools of the organization that created ThinkSpace, so there were pre-existing relationships with the teachers at these schools for many years. In order to keep the identities of the schools, the teachers, and the students private, the schools will be called "Circle" and "Square" for the duration of this paper.

The ThinkSpace curriculum was first implemented when the students were in sixth grade. One teacher from each school opted in to the ThinkSpace study. By opting-in, the teachers were volunteering their students and their classroom to be used by the ThinkSpace program. In total 163 students, about one fourth of all sixth grade students in the district, were taught seasons using ThinkSpace. The remaining students who were not a part of the study learned seasons using the traditional curriculum.

A researcher from the ThinkSpace team took over the two participating classrooms for the duration of the eight-day seasons curriculum. Before any instruction on seasons, each ThinkSpace student took a ten-question pre-test to determine their existing knowledge. The test was a distractor-driven MOSART (Misconceptions-Oriented Standards-Based Assessment Resources for Teachers) test (See Appendix) with questions from Sadler et al.'s concept inventory. The concept inventory is a collection of multiple choice questions based on misconceptions about astronomy and earth science. Researchers from the ThinkSpace team picked the questions they deemed most relevant to the seasons curriculum. Each ThinkSpace student was given an identification number and astronomy word so their names remained anonymous and their pre-test could be matched to their post-test. The students who were not using ThinkSpace did not take the pre-test or the post-test. The students in the study then learned seasons using all of the resources and

assignments built in to the ThinkSpace curriculum. Immediately after they had been taught seasons using ThinkSpace, each ThinkSpace student took a post-test, which was the same as the pre-test, to measure how much they had learned.

Two years later, when the original students in the study entered eighth grade, the school district changed the structure of the science classes so that astronomy would be taught in eighth grade from that point on. One of the main motivations for this is the finding that spatial skills improve with age (Uttal et al. (2013)). Thus, the students who had taken astronomy in sixth grade were required to take it again in eighth grade. This provided an opportunity to test the students again to determine how much of the seasons curriculum they remembered two years after first learning it. The ThinkSpace seasons curriculum was tested in many school districts in Massachusetts in sixth grade, but this specific school district was the only one in which circumstances allowed the ThinkSpace team to test the students after a prolonged period of time.

In this part of the study, all of the eighth-grade students took a pre-test immediately prior to instruction on seasons. The eighth-grade pre-test included the same distractor-driven MOSART questions that were used to test the ThinkSpace students in sixth grade. The eighth-grade pre-test consisted of only eight questions though; two questions on lunar phases were removed for time purposes and because they were previously identified as having low discrimination values. For the statistical analyses that are done in this paper, only the eight questions that are common to both tests are considered. Every student was given an identification number and astronomy word. For the students who had used ThinkSpace in sixth grade, there was no connection between their sixth-grade identification and their eighth-grade identification. For the eighth-grade pre-test, the students who had used ThinkSpace in sixth grade were the experimental group and the students who had not used ThinkSpace in sixth grade were the control group.

All of the eighth-grade students in the two schools used ThinkSpace. There were eight teachers and 586 students total. The eighth-grade teachers will be referred to as letters of the alphabet ranging from A to H. See Table 1 for the break down of teachers and number of students. The setup for the eighth-grade portion of the study differed from the setup for the sixth-grade portion of the study in two ways. First, whereas the two sixth-grade ThinkSpace teachers had chosen to participate in the study, all of the eight grade teachers were mandated by the school district to use ThinkSpace. Second, instead of a researcher from the ThinkSpace team taking over the classrooms, all of the eighth grade teachers were trained how to use the curriculum and were responsible for implementing it themselves. This meant that how the ThinkSpace curriculum was used was left to each teacher's discretion.

Table 1. The eighth-grade teachers from each school, listed under aliases, and the number of students each teacher had. There are two schools with four teachers each and 586 students total.

School	Teacher	No. of students
Circle	A	39
	B	70
	C	80
	D	84
Square	E	80
	F	82
	G	73
	H	78

Immediately after all of the eighth graders learned seasons using the ThinkSpace curriculum as their teachers saw fit, they took the eighth-grade post-test, which contained the the same questions as the eighth-grade pre-test. Six of the teachers administered the test on paper and two of the teachers administered the test electronically. All of the data was assembled into a spreadsheet containing the student's identification number and astronomy word, letter-answers to each of the multiple choice questions, whether they answered each question correctly, their total score, and their gain from eighth-grade pre-test to eighth-grade post-test.

2.2. *Statistical Analysis*

All of the statistical analysis was done in Stata. We performed t-tests and a regression on the eighth-grade data and calculated the effect size for the eighth-grade pre-test.

2.2.1. *T-test*

A t-test compares the means of two groups and determines whether or not a difference between the two means is statistically significant. The null hypothesis assumes the two populations are indistinguishable. The null hypothesis is rejected if the difference between the means of the two groups is statistically significant, which is measured by the p-value. The p-value gives the probability that the difference in the two means occurred by chance. Usually a p-value of 0.05 or less is considered significant. Although $p = 0.05$ is the accepted threshold for statistical significance, there is still a 1 in 20 probability that the difference occurred by chance.

Students who used ThinkSpace in sixth grade will be called "ThinkSpace students" and students who did not use ThinkSpace in sixth grade will be called "non-ThinkSpace" students. A t-test was performed on the eighth-grade pre-test scores to compare the mean score of the ThinkSpace students to the mean score of the non-ThinkSpace students.

2.2.2. *Effect Size*

Whereas a t-test determines if the difference in the two means is statistically significant, the effect size quantifies the difference in the means. Cohen's d, which is one measure of effect size, determines the difference in the means relative to the standard deviation of the data (Equation 1).

$$\text{Effect size} = \frac{\text{ThinkSpace mean} - \text{No ThinkSpace mean}}{\sigma \text{ of eighth grade pretest}} \tag{1}$$

According to Cohen, $d = 0.2$ is a small effect size, $d = 0.5$ is a medium effect size, and $d = 0.8$ is a large effect size (Cohen (2013)).

2.2.3. *Regression*

When there are one or more factors that are suspected could influence the dependent variable, a regression is used to quantify how much influence each factor has. Each independent variable has a corresponding regression coefficient. The regression coefficient describes how steep the regression line is due to that variable. A higher coefficient means the variable has a larger influence on the dependent variable.

In this regression, the dependent variable is the eighth-grade post-test score and the independent variables are the eighth-grade pre-test score, whether or not a student used ThinkSpace in sixth grade, and which eighth grade teacher the student had. The data of students who had teacher H were omitted because the teacher was unable to complete the ThinkSpace curriculum due to medical circumstances.

3. RESULTS

The t-test on the eighth grade pre-test found a significant difference between the scores of the ThinkSpace students and the scores of the non-ThinkSpace students (See Table 2). The mean eighth-grade pre-test score out of eight questions for non-ThinkSpace students was 3.98 ± 0.06 and the mean eighth-grade pre-test score for ThinkSpace students was 4.73 ± 0.13. The p-value was less than 0.0001.

T-tests were done question by question on the eighth-grade pre-test comparing the students who used ThinkSpace in sixth grade to the students who did not. The t-test for each question shows the fraction of students in each group who got the question correct, the standard error, and the p-value. The data is compiled in Figure 8. The main idea of each question in summarized; for the full test questions see Appendix A. A larger fraction of ThinkSpace students got seven out of the eight questions correct. Of those seven questions, the differences between the two groups were

Table 2. A t-test comparing the mean score on the eighth-grade pre-test of the students who used ThinkSpace in sixth grade versus the students who did not use ThinkSpace in sixth grade. The difference is statistically significant.

	Observations	Mean	Standard Error
No ThinkSpace	434	3.98	0.06
ThinkSpace	148	4.73	0.13
Combined	582	4.17	0.06
Difference		-0.76	0.13

$p < 0.0001$

significant for five. In the first question a larger fraction of non-ThinkSpace students got the question correct but the difference is not statistically significant and falls within one standard error. The data is plotted against the national average, taken from Sadler et al. (2010). The national averages for each question are from the ASSCI data. No significance testing was done on the national average data.

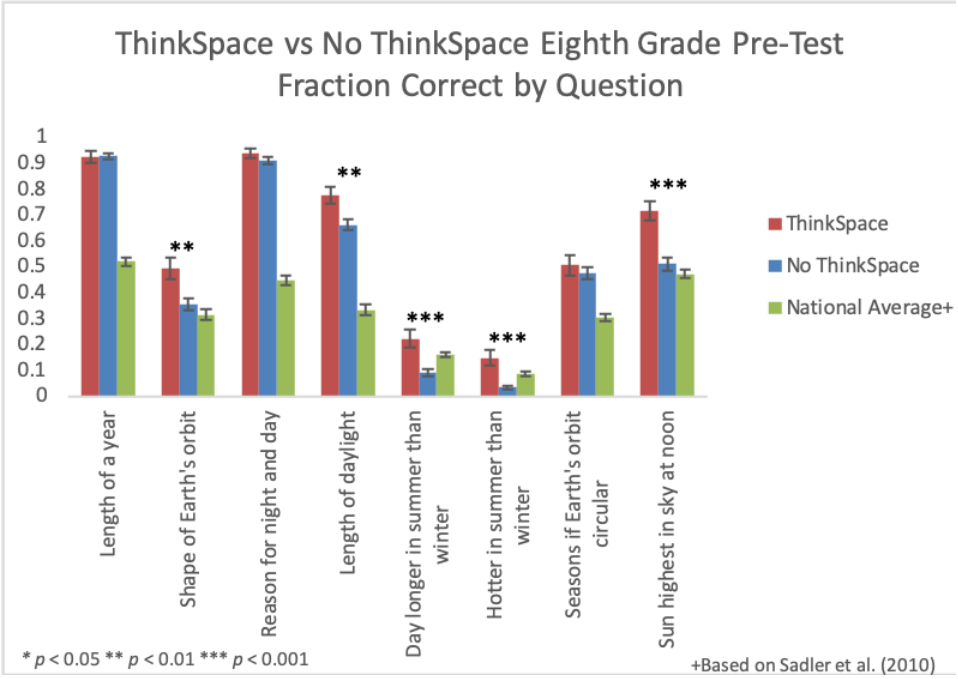


Figure 8. A comparison for each question on the eighth-grade pre-test between the fraction of ThinkSpace students who got the question correct and the fraction of non-ThinkSpace students who got the question correct. For seven of the eight questions, a larger fraction of ThinkSpace students got the question correct. The differences in five of the eight questions are statistically significant. The error bars show the standard error on each measurement. The green bars show the national average for each question, based on Sadler et al. (2010). The p-values shown only compare the ThinkSpace mean to the non-ThinkSpace mean. There was no significance testing done with the national average.

The Cohen's d effect size was calculated comparing the ThinkSpace mean score and the non-ThinkSpace mean score on the eighth-grade pre-test. We found that $d = .54 \pm 0.01$. This means that the mean eighth-grade pre-test score is about 0.54 standard deviations higher for the students who used ThinkSpace in sixth grade compared to the students who did not.

After all of the students used the ThinkSpace curriculum in eighth grade, a linear regression was used to determine how much influence certain factors had on a student's eighth-grade post-test score. The factors that were considered were the student's eighth-grade pre-test score, whether or not the student used ThinkSpace in sixth grade, and which

eighth-grade teacher the student had. Teacher H was unable to complete the curriculum due to medical circumstances, so their data was dropped from all eighth-grade post-test analyses.

In the regression, the only statistically significant factor is the student's eighth-grade pre-test score (See Table 3). The regression coefficient for the eighth-grade pre-test score means that for each additional point a student scores on the eighth-grade pre-test, it is expected that their eighth-grade post-test score will increase by 0.27. After using the ThinkSpace curriculum in eighth grade, whether or not a student used ThinkSpace in sixth grade is insignificant. The regression coefficients for each of the teachers is measured relative to A, which is why the A coefficient is omitted. Which eighth-grade teacher a student had is an insignificant factor for their eighth-grade post-test score. The adjusted R^2 is 0.09. The standard error for each independent variable is given.

Table 3. A linear regression in which the dependent variable is the eighth-grade post-test score. The only significant predictor is the student's score on the eighth-grade pre-test. Whether a student used ThinkSpace in sixth grade is now insignificant. Which teacher the student had is insignificant.

	Post Test	**Standard Error**
Pre-test Score	0.27^{***}	0.04
ThinkSpace	-0.05	0.13
Teacher		
A	0	(.)
B	0.17	0.25
C	0.27	0.25
D	0.11	0.25
E	-0.02	0.25
F	-0.13	0.25
G	-0.28	0.25
Constant	5.65^{***}	0.25
Observations	485	
Adjusted R^2	0.09	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The fraction of students in this study who got each question correct on the eighth-grade pre-test and the eighth-grade post-test can be compared to the national average (from Sadler et al. (2010)). In eighth-grade, every student took a pre-test before seasons instruction. Then all of the students in the study used the ThinkSpace seasons curriculum. It was found that for the eighth-grade post-test, whether or not a student used ThinkSpace or not in sixth grade is no longer relevant. So, all of the students in the study are grouped together when calculating the fraction of students who got each question correct on the eighth-grade post-test. In the following graph, all of the student data is also grouped together when calculating the fraction of students who got each question correct on the eighth-grade pre-test in order to illustrate their progression.

The results from the duration of the study are summarized in Figure 10. The graph shows how the mean score for each group progresses. The mean scores and standard errors for each group from the sixth-grade pre-test to the eighth-grade post-test can be found in Table 4. The students are grouped by which school they attended (Square or Circle) and from there, whether or not they used ThinkSpace in sixth-grade. There are four data points for the students who used ThinkSpace in sixth grade: the sixth-grade pre- and post-tests and the eighth-grade pre- and post-tests. There are only two data points, the eighth-grade pre- and post-tests, for the students who did not use ThinkSpace in sixth grade. This is because data was only collected on these students in the eighth grade; there was no control group in the sixth-grade part of the study. As described using the t-test above, the only significant difference in the data points at any stage is in the eighth-grade pre-test. The students who used ThinkSpace in sixth-grade scored significantly higher on the eighth-grade pre-test than the students who did not use ThinkSpace in sixth-grade.

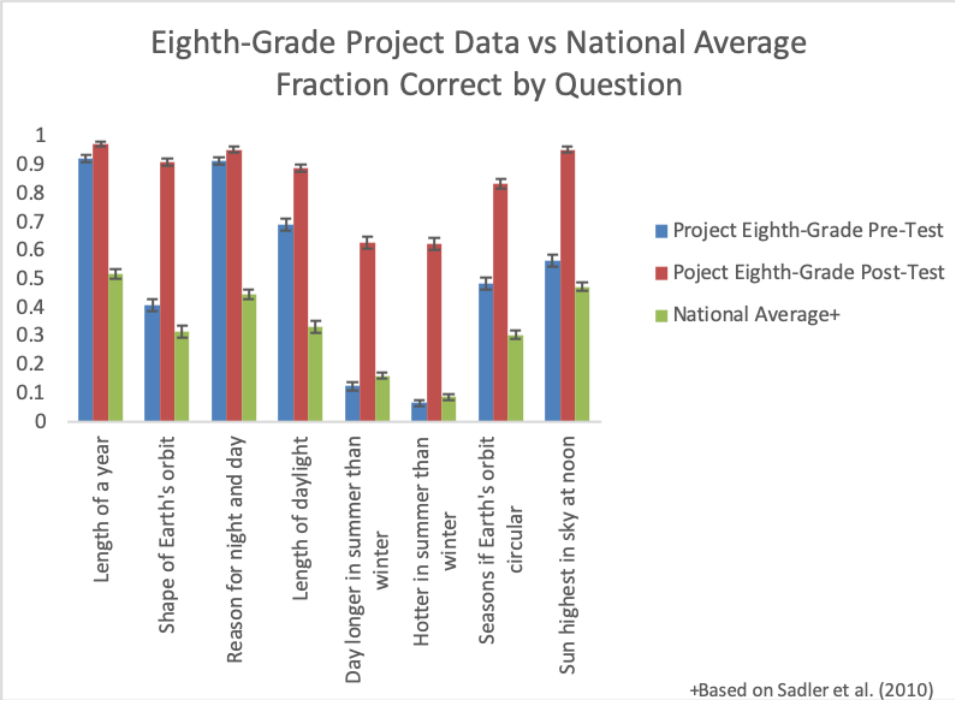


Figure 9. A comparison of the fraction of students who got each question correct on the eighth-grade pre-test, the eighth-grade post-test, and the national average based on Sadler et al. (2010). In this graph, all of the students in the study are grouped together for the eighth-grade pre- and post-tests. The error bars show the standard error on each measurement.

Looking at the data for the students who used ThinkSpace in sixth grade, there is a clear trend in the graph for both schools. The students entered the sixth grade with a mean sixth-grade pre-test score of 3.50 ± 0.11. After using the ThinkSpace curriculum, the mean score increased to 5.94 ± 0.12 on the sixth-grade post-test, which is an average gain of 2.44 ± 0.01. There is then a decrease from the mean score on the sixth-grade post-test to the mean score on the eighth-grade pre-test. The mean eighth-grade pre-test score for ThinkSpace students is 4.73 ± 0.13 and 4.00 ± 0.07 for non-ThinkSpace students. Finally, for all four groups, the mean eighth-grade post-test scores are the highest of the four tests. The ThinkSpace eighth-grade post-test mean is 6.87 ± 0.12 and the non-ThinkSpace eighth-grade post-test mean is 6.75 ± 0.07.

Table 4. The progression of the mean score on each test divided into four groups. The data in this table corresponds to the data shown in Figure 10. The only statistically significant difference is on the eighth-grade pre-test; the mean score of the students who used ThinkSpace in sixth grade is higher than the mean score of the students who did not use ThinkSpace in sixth grade.

	6th Grade Pre		6th Grade Post		8th Grade Pre		8th Grade Post	
Group	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
Square ThinkSpace	3.40	0.19	6.04	0.20	4.60	0.18	6.62	0.19
Circle ThinkSpace	3.59	0.23	5.85	0.22	4.73	0.19	7.06	0.14
Square No ThinkSpace					4.08	0.08	6.64	0.10
Circle No ThinkSpace					3.94	0.10	6.84	0.09
					***			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. DISCUSSION

There was a significant difference in the mean score on the eighth-grade pre-test between the students who used ThinkSpace in sixth grade and the students who did not (See Table 2). The difference between the ThinkSpace stu-

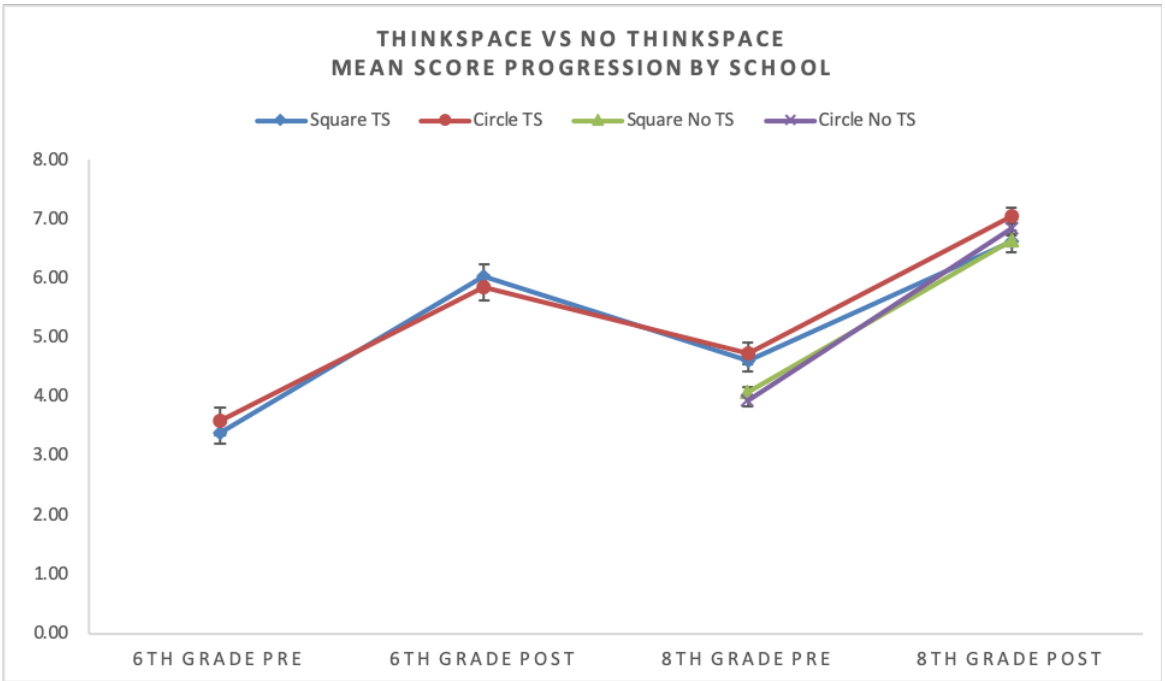


Figure 10. The progression of the mean seasons test score, divided into four groups based on which school the student attended and whether or not they used ThinkSpace in sixth grade. The data for non-ThinkSpace students begins in eighth grade because no data was taken for non-ThinkSpace students in sixth grade. The only significant differences are between the ThinkSpace means and the non-ThinkSpace means on the eighth-grade pre-test. The error bars show the standard error for each mean.

dents and the non-ThinkSpace students suggests that the ThinkSpace curriculum is more effective than a traditional seasons curriculum when measuring how much material students remember after a long period of time. The p-value of the t-test was very low ($p < 0.0001$), and the Cohen's d effect size further confirms this conclusion. We found $d = 0.54 \pm 0.01$. Cohen classifies $d = 0.5$ as a medium effect size, but since the data was taken two years after the treatment, this is a very strong result.

For each question on the eighth-grade pre-test, a higher fraction of students who used ThinkSpace in sixth grade got each question correct compared to the national average (Sadler et al. (2010))(See Figure 8). For six questions on the eighth-grade pre-test, the difference between the fraction of ThinkSpace students who got the question correct and the national average appears to be large (the difference between them is multiple standard errors). This implies that the ThinkSpace curriculum is more effective in the long run, but there were no tests of significance done to compare the ThinkSpace and non-ThinkSpace students to the national average, so it cannot be stated that the differences are significant. It must also be noted that the populations being compared are not the same; the students in this study are from a suburban public school district in Massachusetts and the national average represents students from all over the country from many different backgrounds. Future research on ThinkSpace should include a broader population so that the ThinkSpace data can be more accurately compared to the national averages.

The eighth-grade pre-test was an important part of this study for two reasons. First, it provided an opportunity for the ThinkSpace students to be tested against a control group. Having a control group provides a baseline to compare the results of the intervention to while holding other factors constant. If there is no control group, it could be unclear which factor caused the results. Second, the eighth-grade pre-test was a delayed test from when the students were first taught seasons in the sixth-grade. The students took the sixth-grade post-test immediately after having learned the material, so it was expected that they would recall most of the information. Contrary to this, the eighth-grade pre-test was given two years after they had learned seasons, so it was a measure of how much each student remembered in the long-term based on which curriculum they had used in sixth grade.

On the eighth-grade pre-test, the questions with a statistically significant difference between the ThinkSpace students and the non-ThinkSpace students are believed to require more spatial reasoning skills. On questions one and three of the eighth-grade pre-test, the fraction of ThinkSpace students and the fraction of non-ThinkSpace students who answered the question correctly are within about one standard error of each other (See Appendix A for the questions). Sadler et al. (2010) found that nationally, these two concepts in astronomy are common misconceptions for grades five through eight, so it is likely that the teachers in lower grades in this district did consistently well teaching those particular concepts. This study does not take into account the wording of the questions and how that may affect how many students answer the question correctly. Since the goal of this research is to compare two groups taking the same tests, we assume that a poorly worded question or answer choice will affect both groups similarly.

The linear regression on the eighth-grade post-test score with multiple independent variables yields the following notable results. One independent variable was the eighth-grade pre-test score. The coefficient is positive and it is the only statistically significant factor. This result means that a higher eighth-grade pre-test score predicts a higher eighth-grade post-test score. This is expected because students who have more correct existing knowledge will most likely know more of the material by the end of the course. Another factor that was tested was whether or not a student used ThinkSpace in sixth grade. For the eighth-grade post-test, this factor was insignificant. This is an encouraging result. At the beginning of eighth-grade, there was a significant difference between the students who had used the ThinkSpace curriculum in sixth grade and the students who had not. After all of the students used ThinkSpace in eighth grade, the original difference is no longer present. This means the ThinkSpace curriculum was effective at teaching the seasons material to the eighth-grade students from either background. Lastly, no eighth-grade teacher had a significant impact on the students' eighth-grade post-test scores. This is a positive result because it implies that none of the eighth-grade teachers have a greater influence than the others on the gains seen from the eighth-grade pre-test to the eighth-grade post-test.

The constant in the linear regression is 5.65 ± 0.25. The constant is the predicted score on the eighth-grade post-test if all of the independent variables were zero. The eighth-grade post-test consisted of eight questions, so this is a high constant. The high constant speaks to the strength of the teachers in the district. It should be noted that the R^2 value is 0.09, which is low on a scale that ranges from 0 to 1. The low R^2 value means the regression line is not a very good fit to the data.

From the results of the eighth-grade post-test, it appears that a much higher fraction of the students in this study answer each question correctly compared to the national average. Again, there was no significance testing done on this comparison so nothing can be said with confidence. But based on Figure 9, the difference between the fraction of students in the study who answer each question correctly and the national average is multiple standard errors for each question. The population of the students in this study and the population of the students in the national average is still very different and must be considered when analysing the data. Another reservation is that the students in this study took the eighth-grade post-test immediately after instruction on seasons.

The results from each portion of the study are seen in Figure 10. There was an increase in the mean score from the sixth-grade pre-test to the sixth-grade post test. It cannot be concluded whether or not the ThinkSpace curriculum was the main cause of this gain or if a traditional seasons curriculum would have produced the same result because there was no control group for this portion of the study. For the ThinkSpace students, the mean eighth-grade pre-test is higher than the mean sixth-grade pre-test but lower than the mean sixth-grade post-test. This is expected. Entering the eighth grade, the students should know more than they did before they took an astronomy class in sixth grade. But in the two years between the sixth-grade post-test and the eighth-grade pre-test, most of the students will forget some material. Despite the decrease in mean score, the students who used ThinkSpace in sixth grade did significantly better on the eighth-grade pre-test than the students who did not use ThinkSpace in sixth grade. This was shown by the t-test and the effect size. Finally, all students improved to an equally high level after all using ThinkSpace in eighth grade.

It should be noted that the mean eighth-grade post-test score is higher than the mean sixth-grade post-test score. In most studies, student gains decline significantly when teachers implement the curricula due to poor fidelity of implementation. Fidelity of implementation is ensuring key curriculum procedures are followed when administering

the intervention. The ThinkSpace curriculum is different in this respect because the teacher's outcomes were so strong compared to the team member's results. There are multiple factors that might explain this. One possibility is the teachers have a better relationship with the students than the researchers do, which might have contributed to the students learning more. Another possibility is the students had already learned seasons once before so the material may have made more sense the second time. It is also possible that the students' scores increased because their spatial skills improved from sixth grade to eighth grade. In prior research on ThinkSpace, the spatial skills of students who used ThinkSpace in sixth-grade were also tested. It was found that a student's spatial-skill score is a strong predictor of how well they do on the sixth-grade post-test (Plummer et al. (2020) in preparation). Since it is known that spatial skills improve with age (Uttal et al. (2013)), this could have contributed to the eighth-grade post-test scores in this study being higher than the sixth-grade post-test scores. Since none of these factors were controlled for, it is not possible to say how much, if any, each contributes to the overall higher post-test scores in the eighth grade.

The limitations of this study must be noted. The two schools in this study are in a suburban area of Massachusetts. Since Massachusetts has some of the best schools in the country (Massachusetts was ranked in the top five states for science at the eighth-grade level in 2015 NAEP (2015)), the findings of this research cannot be generalized to say that ThinkSpace will be as effective anywhere else. Furthermore, the implementation of the curriculum was slightly different in the sixth grade and in the eighth grade, and both methods have positive and negative aspects. In the sixth-grade portion, two teachers volunteered their classroom and students. This meant they allowed a researcher from the ThinkSpace team to come in and teach for the duration of the eight-day seasons curriculum. This method eliminates fidelity of implementation issues. The drawback of having trained researchers teach the students is that they do not have the same relationship with the students that the teachers have. A teacher who is more familiar with their students may know a better way to present a topic based on how their students learn best. In the eighth grade, the district teachers administered the curriculum. The reservation with this method is that since the curriculum was mandated by the schools instead of being opt-in, the teachers were not formally a part of the study and could use the curriculum at their own discretion. They were trained how to use ThinkSpace, but there was no enforcement on the implementation. This was not controlled for in the results. The eighth-grade teachers did have the benefit of the existing relationship with their students.

5. CONCLUSION

ThinkSpace is a middle school seasons curriculum which is comprised of activities that are based on spatial thinking. The goal of this study was to determine if students who use ThinkSpace learn more in the long-term than students who use a traditional seasons curriculum. By testing students two years after they had taken an astronomy class, we found a statistically significant difference in the scores of the students who had used ThinkSpace compared to the scores of those who had not. Out of eight questions, the ThinkSpace students answered 0.76 ± 0.13 more questions correct than the non-ThinkSpace students. The effect size for this gain was $d = 0.54 \pm 0.01$. Therefore, in an answer to our motivating research question, we conclude that the spatially-focused ThinkSpace curriculum was more effective than an ordinary seasons curriculum in the long-term. This result could impact the way astronomy and other STEM courses are taught in the future, transitioning to a more spatial approach to learning.

This education research was a design and development study. It began with the idea of improving middle school astronomy curricula by including activities that embed spatial strategies. The curriculum was created based on what methods were thought to have the most impact. Then the curriculum was tested against a control. Since the project was new, testing was only done under limited circumstances; the schools that participated in the study were constrained to Boston and the surrounding area. Based on the positive results of this study, the next step is to test the ThinkSpace curriculum on a broader scale in a more typical setting.

In order to make the results of this paper generalizable to the whole country, the future direction of this project is to test the ThinkSpace seasons curriculum in many more school districts. The study will include urban, suburban, and rural areas across the country; public, private, parochial, and charter schools; and areas with varying socioeconomic environments. 150 teachers with a variety of backgrounds will be recruited to participate in the study. Half of the teachers will use their traditional seasons curriculum and the other half will receive online professional development on how to implement ThinkSpace. Both groups of teachers will administer pre- and post-tests on the seasons material.

The learning outcomes for the two groups would give a more holistic view of whether or not ThinkSpace students do better than a control group in learning seasons.

ACKNOWLEDGMENTS

I would like to thank Dr. Patricia Udomprasert for allowing me to use the ThinkSpace data and for her continued advisement and support throughout this research project. I would also like to thank Dr. Gerhard Sonnert for his detailed comments and feedback on the preliminary draft of this paper. Finally, I would like to thank Professor Xingang Chen for his advice on the presentation of the results and for leading the astronomy research tutorial that this paper was written for.

APPENDIX

A.

The questions on the distractor-driven test that were used for the eighth-grade pre- and post-tests.

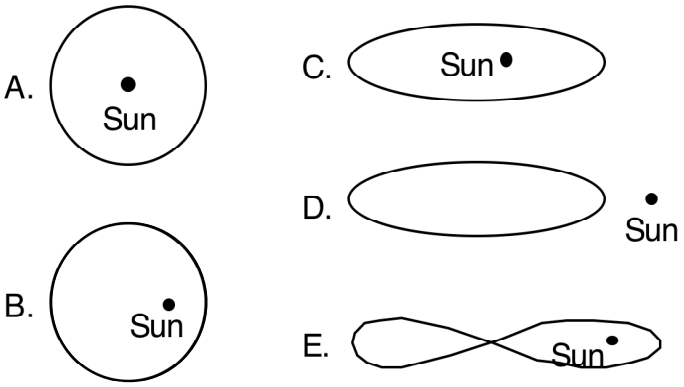
Astro Word: _____	Code # _____	

SEASONS SURVEY

Please answer all questions to the best of your ability. Always choose the single best answer.

1. A year is the amount of time it takes for:
   a. the Earth to rotate once on its axis.
   b. the Moon to orbit the Earth one time.
   c. the Sun to orbit the Earth one time.
   d. the Sun to rotate once on its axis.
   e. the Earth to orbit the Sun one time.

2. Of the following choices, which looks most like the Earth's path around the Sun? *NOTE: Diagrams are the view from far above.*



3. Scientists explain that we have night and day because:
   a. the Sun goes out.
   b. the Earth moves around the Sun.
   c. clouds block out the Sun's light.
   d. the Earth turns on its axis.
   e. the Sun goes around the Earth.

4. The length of daylight would change very little throughout the year if the:
   a. Earth's orbit was perfectly circular.
   b. Earth's orbit was half as big.
   c. Earth's orbit was twice as big.
   d. Earth took half the time to travel around the Sun.
   e. Earth's axis was not tilted.

5. Bruce says that the Sun takes 15 hours to cross the sky each day in June, but only 9 hours in December. Which explanation do you agree with most?
   a. “That’s not possible; the Sun is always up for 12 hours.”
   b. “The Sun must travel faster along its path in the sky in December than in June.”
   c. “The Sun must travel a longer path in the sky in June.”
   d. “The statement makes sense because we’re closer to the Sun in June.”
   e. “The Sun stays overhead much longer in June.”

6. The main reason for it being hotter in summer than in winter is:
   a. the Earth’s distance from the Sun changes.
   b. the Sun is higher in the sky.
   c. the distance between the northern hemisphere and the Sun changes.
   d. ocean currents carry warm water north.
   e. the Sun produces heat and light at a faster rate in the summer.

7. If the Earth’s orbit was perfectly circular, what would happen to the seasons?
   a. There would no longer be seasons.
   b. The southern hemisphere would always be in winter.
   c. The northern hemisphere would always be in winter.
   d. Fall and spring would last twice as long.
   e. We would have basically the same seasons.

8. About what time of year is the Sun highest in the sky at noon?
   a. The first day of Spring
   b. The first day of Summer
   c. The first day of Fall
   d. The first day of Winter
   e. At noon the Sun appears always to be the same height in the sky.

REFERENCES

Cohen, J. 2013, Statistical power analysis for the behavioral sciences (Routledge)

Galano, S., Colantonio, A., Leccia, S., et al. 2018, Physical Review Physics Education Research, 14, 010145

Liben, L. S., & Downs, R. M. 1993, Developmental Psychology, 29, 739

NAEP. 2015, NAEP State Profiles. https://www.nationsreportcard.gov/profiles/stateprofile?chort=2&sub=SCI&sj=&sfj=NP&st=MN&year=2015R3

Ojala, J. 1992, International journal of science education, 14, 191

Palen, S., & Proctor, A. 2006, Astronomy Education Review, 5, 23

Plummer, J., Vaishampayan, A., Udomprasert, P., et al. 2020, in preparation

Plummer, J. D., Bower, C. A., & Liben, L. S. 2016, International Journal of Science Education, 38, 345

Sadler, P. M., Coyle, H., Miller, J. L., et al. 2010, Astronomy Education Review, 8, 010111

Sneider, C., Bar, V., & Kavanagh, C. 2011, Astronomy Education Review, 10

Udomprasert, P., Goodman, A., Ladd, E., et al. 2019, in Astronomy Education, Volume 1, 2514-3433 (IOP Publishing), 9–1 to 9–22. http://dx.doi.org/10.1088/2514-3433/ab2b42ch9

Uttal, D. H., Meadow, N. G., Tipton, E., et al. 2013, Psychological Bulletin, 139, 352

Wai, J., Lubinski, D., & Benbow, C. P. 2009, Journal of Educational Psychology, 101, 817