

## ABSTRACT

SHAH, MEETAL JASWANT. Applying a Validity Argument Framework to Learning Trajectories on Middle Grades Geometric Similarity Using Learning Science and Psychometric Lenses. (Under the direction of Dr. Jere Confrey).

Significant attention is being paid to how assessments aligned to learning trajectories (LTs) can be used as a means to locate students' progress in learning big ideas in mathematics over time. Furthermore, data from LT-aligned assessments have been signaled as key to driving student-centered instruction on a day-to-day basis. For LTs and accompanying classroom assessments to become credible sources of evidence of student learning, it is vital for researchers to study the validity of assessments that claim to measure it. Until recently, validation studies of assessments have been conceptualized in the context of large-scale assessments. However, calls for research into conceptualizing validity in the context of classroom-based assessments tied to learning theories (e.g. learning trajectories) have been numerous. This research study is situated within the context of a larger project that studies the implementation and use of a LT-based digital learning system called *Math Mapper 6-8*. It is a mixed-method study that reports on the conduct of a validation study of a learning trajectory for student reasoning about geometric similarity. Two years of field test data from across two districts were analyzed using an Item Response Theory model. Additionally, think-aloud interviews were conducted with 18 middle grades' students. A dual method of integrating qualitative learning science analyses with quantitative IRT analyses underscore the importance of adopting both lenses to study the validity of assessments that function in close proximity to instruction. In addition to contributing to the dialogue about validating classroom-based assessments, the study also finds important insights into 8<sup>th</sup> grade students' understanding of geometric similarity. These findings have implications for teachers and writers of standards.

Applying a Validity Argument Framework to Learning Trajectories on Middle Grades  
Geometric Similarity Using Learning Science and Psychometric Lenses

by  
Meetal Jaswant Shah

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

Mathematics Education

Raleigh, North Carolina

2018

APPROVED BY:

---

Dr. Jere Confrey  
Committee Chair

---

Dr. Allison McCulloch

---

Dr. Karen Hollebrands

---

Dr. Temple Walkowiak

## BIOGRAPHY

Meetal Shah was born on May 20, 1983 in Nairobi, Kenya. She grew up in Kenya and graduated from Premier Academy in 2001. In 2002, she moved to Sydney, Australia to pursue tertiary education. She attended the University of Sydney for her undergraduate studies. After graduating with a Bachelor of Science in Mathematics in 2004, Meetal attended the University of New South Wales, Sydney. She earned her Postgraduate Diploma in Education in 2005 and began her teaching career at a learning center in Rose Bay as she awaited news on her application to emigrate to Australia.

After receiving Australian permanent residency, Meetal taught at Queenwood School, Mosman from 2007 until 2014. At Queenwood she taught a range of grades (7 – 12) including the High School Certificate Extension 1 Mathematics course. In 2010, Meetal returned to graduate school to pursue a Masters in mathematics education on a part-time basis at Macquarie University, Sydney. In 2014, Meetal agreed to relocate to Raleigh, North Carolina after marrying her husband on the condition that she got accepted as a doctoral student at North Carolina State University. She worked as a graduate research assistant under the tutelage of Dr. Jere Confrey for the Scaling Up Digital Design Studies(SUDDS) research group during her three and half years in the mathematics education program.

Meetal has accepted a post-doctoral position with the SUDDS research group at North Carolina State University in Raleigh. She will begin working there in the summer of 2018.

## ACKNOWLEDGMENTS

I have been incredibly fortunate to have had strong and supportive mentors at critical junctures of my life, all of whom have influenced my personal and professional trajectories in unmistakably positive ways. I will trace my steps backwards to thank all those who deserve a special mention. I want to sincerely thank my advisor, my mentor, my committee chair, my guru (in the truest sense of the Sanskrit word) Dr. Jere Confrey. I have come to know Jere as someone who is relentlessly identifying and finding solutions for critical issues in education. She meaningfully engages fledgling researchers, such as myself, other researchers on our team, software developers, funders, and the larger research community so that we can become coherent and rigorous in finding better learning solutions for our most important research partners - students. Jere's passion to make an impact in the world of mathematics education is a daily reminder to me about why I wanted to join the PhD program at NC State in the first place. I want to thank Jere for providing me a treasure trove of opportunities and an extraordinarily holistic learning experience. Every decision and every graduate student task was purposefully chosen and designed to draw on my strengths and to encourage me to think deeply about the program of study that I had chosen to pursue.

I also want to thank Dr. Karen Hollebrands, Dr. Allison McCulloch, and Dr. Temple Walkowiak for providing me with significant feedback during the proposal phase and encouraging me to think deeply during the dissertation defense phase.

Of course, I couldn't have pulled any of this off had it not been for my fellow SUDDS-ers. I thank Mike Belcher for having lengthy discussions with me and helping me organize my thoughts; Margaret Hennessey for helping with my writing and always cheering me on with an array of emoji-filled messages; Yungjae Kim for all the technical help and also being a fervent

believer in my abilities to pull it all off; Dr. Will McGowan for interesting and pertinent lunchtime conversations that helped keep me focused on the end goal, and Emily Toutkoushian for being really positive, patient, and always helpful as I tried to figure out relevant aspects of item response theory.

I want to thank the cohort of graduate students at the college of education who started the program in the 2013-15 academic years. There are too many to name individually, but I want to especially thank Derek Williams for providing me with thoughtful and timely feedback on all of my chapters.

I would also like to thank my professional mentors and dear friends: Yvette Semler, Dr. Lloyd Dawe (University of Sydney), and Dr. Michael Cavanagh (Macquarie University). If it were not for them, I would not have had the courage to set foot in the field to begin with. I thank them for their conviction and their tireless encouragement as I tried to figure out what my place in mathematics education was going to be.

At the time of writing this, several hundred children are being separated from their families by the U.S. administration. I shudder to think what kinds of lives these children are going to embark on without the support of their relatives. Thus, I am doubly grateful that I had not just one but two sets of parents. Being so far away from them (by my own design) has arguably been the toughest part of the last three and a half years. However, their life lessons have not been in vain. I want to thank my aunt Anjana and uncle Narendra for teaching me the virtues of hard work and selfless dedication. My upbringing, for which they are responsible, has been the bedrock of all of the successes I have come to enjoy as an adult. My cousin Minal, the sister I never had, has been constantly cheering me on and helping me figure out how to navigate the often-suffocating path that women walk when they choose to pursue their own desires.

Lastly, I want to thank the one person without whom, none of this would have come to fruition. I want to thank my best friend, partner-in-crime, fellow foodie, and the chairperson of the “Meetal fan club” – Mihir.

## TABLE OF CONTENTS

<b>LIST OF TABLES .....</b>	ix
<b>LIST OF FIGURES .....</b>	x
<b>Chapter 1: Introduction.....</b>	1
Background.....	1
Statement of the Problem .....	2
Purpose of Study.....	3
Overall Research Question .....	5
Plan of Thesis .....	6
<b>Chapter 2: Literature Review.....</b>	7
Students' understanding of geometrical similarity .....	7
Learning trajectories.....	13
<i>Math-Mapper 6-8: A middle grades digital learning system</i> .....	20
Assessment Validity.....	24
Measuring and validating LT-aligned assessments.....	27
Setting up the Study .....	32
<b>Chapter 3: Methodology.....</b>	33
Research Questions .....	34
Materials and Tools.....	35
Math-Mapper 6-8 .....	35
The <i>Similarity and Scaling</i> LT in <i>Math-Mapper 6-8</i> . ....	37
Method .....	43
Sample.....	47

Data Sources and Data Analysis Procedures.....	49
Research question 1.....	50
Research question 2.....	54
Research question 3.....	59
Research question 4.....	59
<b>Chapter 4: Results and Discussion .....</b>	<b>62</b>
Results from Think-Aloud Interviews .....	62
Research Question 1.....	62
Summary of Item Analyses from Think-Alouds. ....	106
Summary of Student Understanding of Geometric Similarity from Think-Alouds.....	108
Summary of Raw Scores from Think Alouds.....	110
Patterns of Item Difficulty Statistics within the <i>Similarity and Scaling</i> LT .....	111
Research Question 2.....	111
Comparison of Fit Statistics.....	112
Results from IRT Rasch model.....	113
Distribution of Scores Across Gender and Ethnic Subgroups.....	125
Research Question 3 .....	125
Correlation of Scores from <i>Similarity and Scaling</i> Pretest and an Independent Assessment .	130
Research Question 4.....	130
<b>Chapter 5: Discussion and Conclusion.....</b>	<b>134</b>
Overview of Study .....	137
Student learning of middle grades geometric similarity .....	137
Validating learning trajectories-based assessments .....	140

Limitations.....	142
Closing Remarks .....	143
<b>References .....</b>	<b>144</b>
<b>APPENDICES.....</b>	<b>158</b>
<b>Appendix A: IRB Approval Letter.....</b>	<b>159</b>
<b>Appendix B: Recruitment Letter and Parent Consent Forms .....</b>	<b>160</b>
<b>Appendix C: Example of a Teacher Report for One Construct.....</b>	<b>164</b>
<b>Appendix D: Example of a Student Report .....</b>	<b>165</b>
<b>Appendix E: Big Ideas in Math-Mapper 6-8.....</b>	<b>166</b>
<b>Appendix F: Assessment Items aligned to Similarity and Scaling LT .....</b>	<b>167</b>
<b>Appendix G: Interview Protocol .....</b>	<b>190</b>
<b>Appendix H: Think-aloud Distribution of Items across Student .....</b>	<b>193</b>
<b>Appendix I: Independent Assessment of Geometric Similarity .....</b>	<b>196</b>

## LIST OF TABLES

Table 1. Demographics for All Sites.....	48
Table 2. Demographics for Think-Aloud Interviewees. ....	49
Table 3. Mapping of Data Sources to Research Questions. ....	50
Table 4. Item associated with each level of the <i>Similarity and Scaling</i> LT.....	55
Table 5. Mapping of Independent Assessment Item Bank to <i>MM 6-8</i> LT progress level. ....	61
Table 6. Classification of items based on data from think-alouds.....	108
Table 7. Comparison of Fit Statistics of Two Confirmatory IRT Models.....	112
Table 8. Regression Equations, R <sup>2</sup> , and Sum of Squares for iterative regression models.....	118
Table 9. Comparison of Recommendations from Think-alouds and Regression.....	124
Table 10. Output for ANOVA between gender subgroups.....	127
Table 11. Descriptive statistics of student performance on <i>MM</i> assessments.....	127
Table 12. Output for ANOVA among ethnic subgroups. ....	129
Table 13. Descriptive Summary of Scores from the pretest and independent assessment. ....	131

## LIST OF FIGURES

Figure 1. Complex shapes in scaling tasks taken from Cox (2013, p.10).....	10
Figure 2. Example of a student’s exploration with “growing” a triangle. ....	11
Figure 3. The conceptual corridor (Confrey, 2006).....	13
Figure 4. Battista’s (2004) LT taken from Lobato and Walters (2017, p.77). .....	16
Figure 5. Functions Learning Progression taken from Wilmot et al. (2011, p. 265).....	18
Figure 6. The assessment triangle taken from Pellegrino et al. (2001, p. 44). ....	27
Figure 7. Abbreviated constructs maps taken from Lehrer et al. (2014, p. 43).....	29
Figure 8. Math-Mapper’s learning map. ....	36
Figure 9. Hierarchy of elements that form Math-Mapper’s learning map.....	36
Figure 10. The two LTs associated with Investigating Transformations and Scale.....	38
Figure 11. Assessment item aligned to Level 1 of the Similarity and Scaling LT.....	39
Figure 12. Assessment item aligned to Level 4 of the Similarity and Scaling LT.....	40
Figure 13. Assessment item (item 770) aligned to Level 7 of the <i>Similarity and Scaling</i> LT.....	41
Figure 14. Example of a ratio box representation used for solving item 770. ....	42
Figure 15. Assessment item aligned to Level 9 of the Similarity and Scaling LT.....	43
Figure 16. Convergent design: integration at the interpretation and reporting level.....	46
Figure 17. Groups of students who participated in the study.....	47
Figure 18. Procedure for coding think-aloud interview data for each item .....	54
Figure 19. Item 749, Level 1.....	64
Figure 20. Item 1181, Level 1.....	66
Figure 21. Item 759, Level 2.....	69
Figure 22. Item 762, Level 2.....	71

Figure 23. Item 765, Level 3.....	74
Figure 24. Item 766, Level 3.....	76
Figure 25. Item 767, Level 3.....	77
Figure 26. Item 763, Level 4.....	80
Figure 27. Item 764, Level 4.....	81
Figure 28. Item 768, Level 4.....	82
Figure 29. An example of a student's incomplete drawing used to answer item 768.....	83
Figure 30. TA 6's drawing of the hamster cage.....	84
Figure 31. Item 760, Level 5.....	85
Figure 32. Item 769, Level 5.....	86
Figure 33. Item 750, Level 6.....	89
Figure 34. Item 771, Level 6.....	91
Figure 35. Item 773, Level 6.....	92
Figure 36. Item 800, Level 6.....	93
Figure 37. Item 751, Level 7.....	95
Figure 38. Item 770, Level 7.....	97
Figure 39. Item 236, Level 8.....	99
Figure 40. Item 241, Level 8.....	100
Figure 41. Item 433, Level 8.....	101
Figure 42. Item 761, Level 9.....	103
Figure 43. Item 1182, Level 9.....	104
Figure 44. Current misconception associated with Level 1 of the Similarity and Scaling LT ..	109
Figure 45. Overview of students' raw scores on items they did the think-aloud on.....	111

Figure 46. IRT (Rasch) $\theta$ -parameters by LT level for 21 field tested items.....	114
Figure 47. IRT (Rasch) $\theta$ -parameters by LT level for 21 field tested items.....	115
Figure 48. The baseline linear regression model applied to the subset of 19 items.....	116
Figure 49. Results of the sequential regression models.....	117
Figure 50. Venn diagram showing the integrated results from two methods.....	123
Figure 51. Distribution of theta-scores for female students .....	126
Figure 52. Distribution of theta-scores for male students.....	126
Figure 53. Distribution of theta-scores for White students.....	128
Figure 54. Distribution of theta-scores for Hispanic students*......	128
Figure 55. Distribution of theta-scores for African-American students.....	128
Figure 56. Distribution of theta-scores for Asian students.....	129
Figure 57. Distribution of theta-scores for Mixed Race students.....	129
Figure 58. $\theta$ scores for pretest vs. $\theta$ scores for independent assessment ( $r= 0.42$ , $p < .01$ ).....	131

## **Chapter 1: Introduction**

### **Background**

Research on how school-aged students learn mathematics has accumulated steadily since the inception of scholars' efforts in both the learning sciences and mathematics education (Cai, 2017; English, 2008, 2015; Grouws, 1992; Lester, 2002). The body of research has reached a point where some of it can be synthesized into learning trajectories (Clements & Sarama, 2004; Confrey, Maloney, & Nguyen, 2014; Lobato & Walters, 2017). Learning trajectories are "research-based descriptions of how students' thinking evolves over time from informal ideas to increasingly complex understandings" (Confrey, Maloney, Nguyen, Mojica, & Myers, 2009, p. 3). Because learning trajectories are based on empirical research on student learning, teachers can utilize them to strengthen their pedagogical content knowledge. Indeed, studies have shown that teachers' understanding of LTs can help them become more aware of and thus leverage a variety of student thinking in relation to a mathematical concept when making instructional decisions such as planning lessons, orchestrating discourse, and analyzing student work (Clements, Sarama, Spitler, Lange & Wolfe, 2011; Wilson, Mojica & Confrey, 2013; Wilson, Sztajn, Edgington & Confrey, 2014).

Some of the ways that researchers have sought to help teachers adopt LTs into their instructional practice is by providing them with curricula that are explicitly tied to LTs (Clements & Sarama, 2009; 2014). Others have supported teachers' understanding of LTs through professional development centered around LTs (Sztajn, Confrey, Wilson & Edgington, 2012). One of the persisting concerns around these two models of disseminating research on LTs is that they do not enable teachers to assess, on an ongoing basis, whether they are supporting all of their students (Confrey & Maloney, 2012). One way of getting around this concern is to

provide teachers with assessments that measure students' progress along LTs (Daro, Mosher & Corcoran, 2011). Such assessments fall into a new genre of classroom-based assessments that are undergirded by LTs, administered in close proximity to instruction, and used diagnostically to drive instructional decision making (Confrey & Maloney, 2012).

Confrey (2015) designed a digital learning system (DLS) as a means to address the aforementioned concern about assessing student learning along LTs to ensure that teachers are meeting the needs of all their students. A DLS is “a single platform into which learning targets, learning opportunities provided by digital curriculum materials and tools, assessment activities, and analytic capacity are all situated” (Confrey, Gianopoulos, McGowan, Shah & Belcher, 2017, p. 719). The current study is embedded within a long-term research project that studies the design, implementation, and effects of a DLS called *Math-Mapper 6-8 (MM6-8)*. Confrey (in review) defines the purposes of this DLS as: 1) to inform teachers of class progress by LT, 2) to elicit and leverage diverse levels of student thinking, 3) to increase students' awareness of their own learning and growth, 4) to strengthen teachers' content and pedagogical knowledge, 5) to connect instructionally proximal data to more distal forms of assessment (interim and high stakes), and 6) to improve students' knowledge of target constructs and big ideas from the learning map.

### **Statement of the Problem**

In order to improve outcomes for most or all students in a classroom, it is imperative for teachers to drive instructional decisions based on valid and reliable data about their students that is generated in close proximity to instruction. As it stands, there is little consensus among researchers developing assessments that measure LTs about how to approach the validity of these assessments (Duschl, Maeng, & Sezen, 2011). Furthermore, concepts of validity and

reliability that are developed in the context of large scale assessments. The technical accuracy and stringent benchmarks for reliability and validity developed for large scale assessments are developed with high stakes decision making and consequences in mind (Pellegrino, Chudowsky, Glaser, 2001). However, these need to be reconsidered so they are more suited to the purposes of classroom-based assessments that are tied to learning theories such as LTs (Brookhart, 2003; Pellegrino, DiBello, Goldman, 2016).

### **Purpose of Study**

This study focuses on the diagnostic assessment of one LT within the MM6-8 DLS. Such a diagnostic assessment has the potential of becoming a key component in assessing student understanding within a teacher's daily instructional sequence (Confrey, Hasse, Maloney, Nguyen, & Varela, 2010; National Research Council, 2003). These digitally-administered and scored systems represent a significant subset of activity within a new movement called "classroom assessment" in contrast to high stakes, interim, or informal formative assessment (Pellegrino, Chudowsky & Glaser, 2001; Pellegrino, DiBello, & Goldman, 2016). Classroom assessment occurs in the regular course of instruction because it not only helps teachers to identify gaps in their students' understanding of a given concept, but it also helps them to modify their subsequent instructional decisions (National Research Council, 2003). However, assessments that are proximal to instruction need to be designed so that they allow for productive, data-driven instructional decisions if they are to serve their intended purpose (National Research Council, 2003; Pellegrino, DiBello, & Goldman, 2016). In contrast to large-scale, high stakes assessments, classroom assessments must be frequent (but not obtrusive), assess proficiencies that are not easily measured by large-scale assessments, provide quality feedback to both teachers and students, flag students' misconceptions, and enable students to

capitalize on the resultant feedback (National Research Council, 2003). All assessment formats need careful consideration regarding the validity of the interpretations derived from their score results (National Research Council, 2003; Pellegrino et al., 2016). However, the literature on assessment validity has focused primarily on large-scale, high stakes assessment (Pellegrino et al., 2016). Thus, this study seeks to pioneer the conduct of such studies in the context of classroom assessment and contribute to a) the current discourse about developing a measurement theory that is specific to the purposes and uses of classroom assessment, and b) the development of a *lingua franca*, that both learning scientists and psychometricians negotiate in order to conceptualize the validity of such assessments. This shared language can help both parties describe the priorities and purposes of classroom assessments which are in many ways unique to their settings. Furthermore, it can help delineate multiple, relevant sources of evidence that will be integrated in a principled manner (Nichols, Kobrin, Lai & Koepfler, 2016) to backup claims related to the purposes of the assessments.

Specifically, this study investigates the validity of an assessment of a LT developed for students in the middle grades who are studying geometric similarity. Similarity was chosen because of the paucity of research on how student thinking progresses in this area of geometric reasoning, and because of the topic's strong ties to the notion of ratio, which has been a central focus of the SUDDS program of work. Similarity was also chosen because outside of school mathematics, the notion of two objects being the same but different at the same time is ever-present in children's daily lived experiences and these can be leveraged extensively during instruction (e.g. responsive web pages that re-size content to fit different device screens, 3D printing, virtual and augmented reality platforms, Computer Aided Design (CAD) animations, etc.) Today's middle grades students live in an era where digital technologies with underlying

core principles of similarity and scale can be used to model airplanes, cars, houses, as well as devices that end up in space. Although many of these middle grades students may not go on to pursue an Erlanger-like program in geometry as defined by Felix Klein, the notion of invariant properties is a powerful tool that will continue to help them understand the world around them.

Thus, this dissertation study applies the SUDDS validation argument to the case of an LT-aligned assessment addressing middle grades geometric similarity content. The study uses a variety of methods to gather and integrate evidence to support four specific claims about the assessment which are addressed by the research questions listed below.

### **Overall Research Question**

To what extent do the assessments aligned to the Similarity and Scaling learning trajectory within the digital learning system Math-Mapper 6-8 provide a user with valid information about student thinking along a continuum of learning?

1. To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in MM 6-8?
2. To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?
3. How do patterns of observed performance differ across students from a diverse sample on LT-based assessments measuring geometric similarity?
4. How do the students' results on the MM 6-8 Similarity and Scaling assessment compare to those on an independent assessment measuring geometric similarity?

These four questions address a subset of claims made in the full validation argument for *MM 6-8*'s diagnostic assessments and they provide substantial information on key elements of the SUDDS team's validation approach.

### **Plan of Thesis**

Following this introductory chapter, in Chapter Two, a synthesis of the literature justifying the need for this study is presented. The synthesis will be divided into sections that review the research on student learning in geometric similarity as well as the literature on learning trajectories/progressions. These two sections will illustrate how *the Scaling and Similarity* LT and the assessment items that measure the levels draw on these two bodies of research. Next, the body of research addressing assessment validity in general, as well as validity in the context of LT-aligned assessments, will be reviewed so as to set the stage for the specific validity claims that will be addressed by this study.

Chapter Three provides a brief background of the SUDDS project and its validity argument first so as to position the remainder of the thesis. The background is followed by details about the research participants and specific methodologies applied to the data collection and data analyses processes. Chapter Four is dedicated to the results of the study organized into four sections, one for each of the four research questions. Each results section necessitates a discussion of the findings and an overall discussion is used to summarize the chapter. The final chapter includes an overview of the study and a synthesis of the findings followed by an outline of the future directions, implications of the findings and concluding remarks.

## **Chapter 2: Literature Review**

Validation of a learning trajectory on similarity requires a synthesis of literature from mathematics education, the learning sciences, and assessment. Too often these fields function at a distance from each other and arguably the quality of assessment can suffer from insufficient integration (Pellegrino et al., 2001). The chapter begins with the research on student learning in geometric similarity followed by a review of the literature on learning trajectories/progressions. These two sections will illustrate how the Scaling and Similarity LT and the assessment items that measure the levels in the tool, *Math-Mapper 6-8*, draw on these two bodies of research. Next, the body of research addressing assessment validity in general, as well as validity in the context of LT-aligned assessments, will be reviewed so as to set the stage for the specific validity claims that will be addressed by this study. A section is included describing the background of the SUDDS project and its current validity framework that serve as the context and foundation of this work.

### **Students' understanding of geometrical similarity**

Anecdotally, young children have several experiences with similarity in context. For example, they often play with model airplanes, cars, and houses to recognize that objects can be the “same” (same shape) and “not the same” (different size) all at once. Children who are old enough to work with mobile and desktop computers encounter responsive webpages that re-size the same content to fit different device screens. In more high-tech environments, children are also seen to create scale-based models of a variety of objects by employing 3D printing and Computer Aided Design (CAD) and animation software at local libraries, museums, and maker spaces.

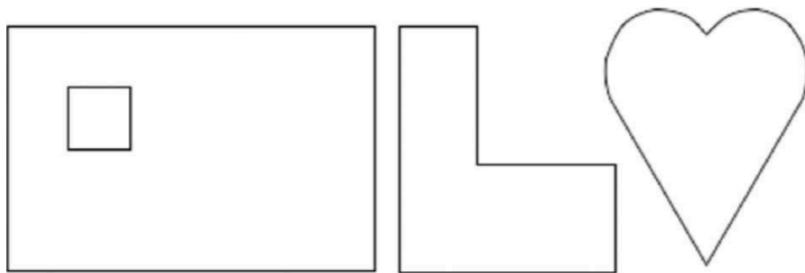
In research, studies have shown that students coming to middle school already have an intuitive understanding of similarity (Jones, 2000; Lehrer, Strom & Confrey, 2002). It is important for teachers to recognize that students' everyday notions of similarity can be leveraged during instruction, but that students may also come to class with common misconceptions. For example, students may believe that two objects to be categorized as being the "same," they must be oriented in the same way (Lehrer, Jenkins & Osana, 1998). Researchers attribute this misconception to students' regular encounters with prototypical shapes before they formally study geometrical objects (Clements & Battista, 1992; Vinner & Hershkowitz, 1983;). Clements and Battista (1992) state that even middle school students have persisting difficulties with identifying shapes that are not presented in prototypical manner. Panorkou, Maloney, Confrey and Platt (2014) suggest that such enduring misconceptions may be an artifact of the static images that continue to be employed in instructional materials despite the availability of several dynamic geometry environments (e.g. Sketchpad and GeoGebra).

Leveraging intuitions about similarity as a springboard into a more formal treatment of geometric similarity requires teachers to draw students' attention to the properties of shapes and figures (Confrey, 1992; Cox & Lo, 2014; Lehrer et al., 2002) using appropriate tasks. For example, a study conducted by Confrey (1992) with elementary grade students gives insight into a range of informal reasoning approaches that students tend to use when sorting physical cut-outs of shapes (different-sized circles, squares, triangles, and rectangles) into piles of similar shapes. Confrey (1992) notes that students typically use three methods to discern similarity: a) stacking and centering to look for parallel sides, b) stacking and matching angles by sliding one shape over other, and c) trying to match shape in relation to the extent of deviation from a prototypic shape (often an equilateral version of the shape or one with certain symmetries). She established

that first graders could most easily identify similarity in circles and squares, followed by triangles (equilateral, non-isosceles right-angle, and non-right angle isosceles). For these students, rectangles were the most difficult to establish similarity across due to the fact that they all have the same right angles but may not have sides in proportion. When working with tasks that involve comparing physical properties of shapes, it is important for teachers to be aware that the visual strategies as identified by Confrey (1992) that involve checking for congruence of angles at the vertices of shapes have certain limitations. For example, students may incorrectly extend what they notice to be a suitable criterion for judging the similarity of triangles (congruent angles at the vertices) to other shapes like rectangles. Likewise, Chazan (1988) notes that students can sometimes have the misconception that two rectangles are similar because the angles at the vertices are all right angles. Having said that, there are strategies that students can employ in order to visually inspect whether two or more rectangles are indeed similar. Lehrer et al. (2002) in their study find that students can draw several rectangles on a coordinate plane (longs side along one axis and short sides along the other) which allows them to compare the diagonals of the rectangles. Students then establish that two rectangles are similar if a straight line can be drawn along their diagonals.

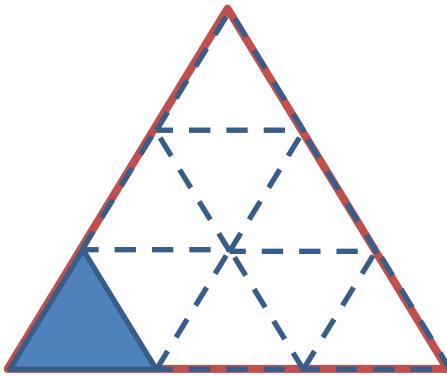
In other research, Lo, Cox, and Mingus (2013) suggest that students should also be given the opportunity to work on tasks that involve informal scaling of uncommon shapes to develop a more robust understanding of similarity. For example, a study done by Cox (2013) shows how middle grades students tasked with informally scaling complex shapes (Figure 1) are encouraged to attend to the quantifiable features of the shape which may not be physically represented by lines (e.g. space around the square that is embedded in the rectangle). Cox (2013) notes that students often use a combination of visual and numerical reasoning when informally scaling such

shapes. For example, after sketching on paper an enlarged version of a shape, students visually inspect their resulting figure to check for any distortions.



*Figure 1.* Complex shapes in scaling tasks taken from Cox (2013, p.10).

According to Cox (2013), “providing students with complex figures to scale may encourage students to mathematize their visual perceptions and increase their ability to attend to the quantifiable features of shape and the numeric relationships between them” (p. 22). This is a productive observation and one that can be further leveraged by teachers, so students can start paying attention to the relationships between side lengths, among other quantifiable attributes, of similar shapes. Teachers can give students tasks that require them to enlarge shapes by doubling or partitioning shapes (Pothier & Sawada, 1990) into smaller congruent parts. Such tasks enable students to converge towards a fuller definition of similarity that entails not just paying attention to angles, but also the side lengths. In the example shown in Figure 2, a student uses iterations of the smaller blue triangle to build the larger red triangle, which helps them see that the larger triangle tends to have the same angles at the vertices and has side lengths that are  $k$  times as long (Confrey, unpublished).



*Figure 2.* Example of a student’s exploration with “growing” a triangle.

As students work with a variety of aforementioned enlarging or equipartitioning tasks, Confrey and Carrejo (2005) recommend equipping students with representations such as ratio boxes and tables so as to draw their attention to the multiplicative relationships between the corresponding side lengths of the triangles. The idea of stretching and shrinking from the literature on rational numbers shows us that the implicit notion of keeping ratios of side lengths the same can also be developed by using “free interaction” activities such as drawing that evolve into activities where students look for “rules for ratios” (Dienes, 2001). As students carry out a variety of stretches (multiply by a scale factor e.g.,  $a$ ) and partitions (divide by a scale factor e.g.,  $b$ ) and continue to document their work using ratio tables they may be able to flexibly scale up or scale down triangles by using each type of dilation explicitly or a combination of both dilations (scale factor e.g.  $\frac{a}{b}$  ).

Once students are using ratio reasoning to establish similarity across triangles, they may be ready to extend their understanding of similarity to other shapes, and use similarity, proportionality, and scale factors to solve problems involving scale drawings such as those that are typically found in textbooks and standardized assessments (French, 2004). Whilst exploring similar shapes, it is important for teachers to encourage students to study the impact that a scale factor of  $k$  has on other measures of the shape such as area for two dimensional shapes and

volume for three-dimensional objects (Lo, Cox, Mingus, 2006). Students can begin by comparing the shapes visually e.g. using the smaller shape to tile the larger shape in order to see “how many” smaller shapes fit inside the larger shape (Confrey, personal communication). Students can also find the area (or volume) of dilated shapes and compare them to recognize that they scale by a factor of  $k^2$ (or  $k^3$ ) (French, 2004).

Overall, the research on student learning of similarity, suggests it is advisable that before students are expected to use ratio reasoning to study geometric similarity, they be given the opportunity to visually inspect shapes and develop their own strategies for distinguishing similar and non-similar shapes and for relating similar shapes to each other. Students can employ a variety of visual inspection strategies to establish similarity across shapes (Confrey, 1992; Cox, 2013). Others have suggested teaching similarity through transformations for example, Usiskin’s original ideas (1972) on teaching similarity and congruence from transformation-based approaches. Usiskin’s work has been rekindled by Seago et al. (2014) which is perhaps another method for helping students connect ratio reasoning to conceptions of similarity. The Common Core State Standards in Mathematics (CCSS-M) also recommend a transformational approach to teaching similarity, however the standards for similarity and transformations are split across two grade levels. 7<sup>th</sup> grades geometry standards address the concept of similarity, whereas 8<sup>th</sup> grade geometry standards address the concept of transformations.

As students establish similarity across shapes using visual strategies, they begin to attend to a variety of measures such as angles and lengths of the shapes, and develop conjectures of what it means for given shapes to be similar. As students enumerate the relationships between the geometrical properties they may have a firmer foundation to be able to work with a variety of

problems that may involve using scale factors e.g. finding missing side lengths or relating given shapes by finding a constant of proportionality.

## Learning trajectories

Learning trajectories (LTs) in mathematics education have advanced Simon's (1995) notion of the "hypothetical learning trajectory" (HLT) and are being developed as tools that document progressions of student understanding of specific mathematical concepts based on empirical research on student learning. Confrey et al. (2009) define LTs as:

a researcher-conjectured, empirically supported description of the ordered network of constructs a student encounters through instruction (i.e., activities, tasks, tools, forms of interaction and methods of evaluation), in order to move from informal ideas, through successive refinements of representation, articulation, and reflection, towards increasingly complex concepts over time. (p. 347).

Confrey's (2006) (Figure 3) *conceptual corridor* better illustrates the idea of this movement from naïve to sophisticated understanding as students encounter obstacles (e.g. misconceptions) and landmarks (e.g. important cognitive ideas).

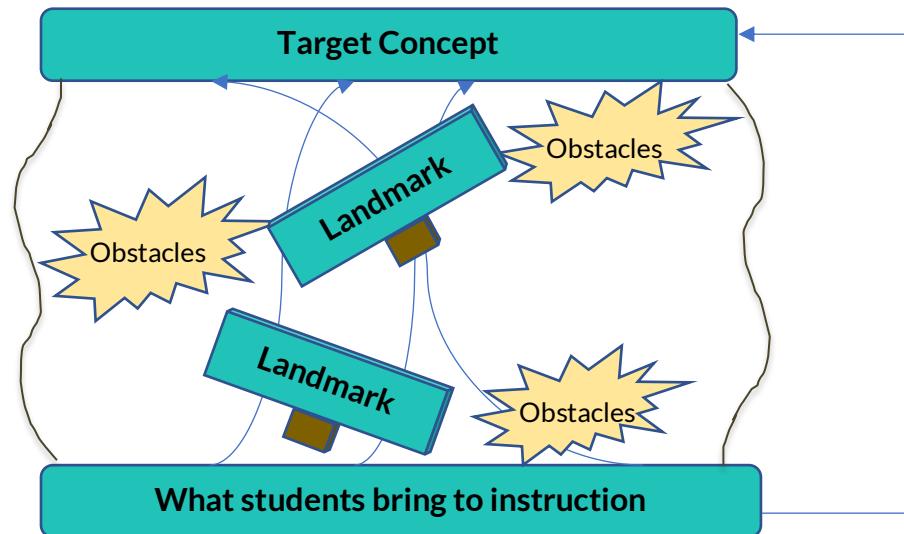


Figure 3. The conceptual corridor (Confrey, 2006).

It is important to clarify at this point that even though LTs delineate pathways towards sophisticated understandings of a mathematical concepts, they are not a stage theory (Confrey,

Maloney & Nguyen, 2014; Lehrer & Schauble, 2015). That is, not all students pass through every single landmark concept or misconception as described in an LT in a lock-step manner. Furthermore, LTs are not a rational analysis of the domain (Clements & Sarama, 2004, Confrey et al., 2014) as can be the case with curriculum documents such as textbooks, standards documents, and web-based resources. In other words, the latter provide teachers with a description of the target concept and a summary of skills that experts in the domain espouse but they do not define possible intermediate phases that are unique to a student's construction of that concept (Confrey et al., 2014).

Similar to the concept of LTs is that of learning progressions (LPs), in that LPs also function as research-based descriptions of student thinking (Duschl, Schweingruber & Shouse, 2007). Both LPs and LTs are geared toward helping teachers shift the focus of their instruction from getting students to solely accumulate facts and skills, to helping students develop domain specific cognitive and meta-cognitive practices as they cultivate knowledge (Duschl et al., 2007; Duschl, Maeng & Sezen, 2011; Lehrer, Kim, Ayers & Wilson, 2014; Pellegrino et al., 2001). Historically, such efforts to articulate empirical research on student thinking in mathematics education have been grouped under the banner titled *learning trajectories* (Clements & Sarama, 2004) whereas parallel efforts in science education have been titled *learning progressions* (Duschl et al., 2007; 2011).

Growing research efforts on LTs (and LPs) have seen them become prevalent not just in the classroom and among researchers, but also among writers of standards. For example, many College and Career Readiness Standards, especially those derived from the Common Core State Standards for Mathematics (CCSS-M), claim a foundation in learning progressions/trajectories (Common Core Standards Writing Team, 2016; Confrey, Maloney, & Corley, 2014; Daro,

Mosher, & Corcoran 2011; Hess & Kearns, 2010). Another example of the prevalence of LTs can be seen in Confrey's (2015) research efforts to develop a LT-based learning system called *Math-Mapper 6-8 (MM 6-8)*.

Despite a growing number of researchers studying LTs (e.g., Battista, 2011; Clements & Sarama, 2007; Confrey et al., 2014; Ellis, Ozgur, Kulow, Dogan & Amidon, 2016; Lehrer et al., 2014) and the education community viewing them as valuable instructional tools (e.g., Corcoran et al., 2009; Daro et al., 2011), there is no universally agreed upon definition of a LT (Lobato & Walters, 2017). Individual scholars differ in the ways that they define the scope of the LT, the grain size of the levels that make up the LT, and the audience of the LT (Battista, 2011; Corcoran et al., 2009; Duschl et al., 2011). In the succeeding paragraphs, a deeper review of specific studies illustrates how LTs in mathematics education have some commonalities but also vary along many dimensions.

LTs typically address fundamental concepts in the domain: they illuminate the naïve conceptions that students bring to instruction in relation to a target concept within a big idea, and they have intermediate descriptions of how student reasoning develops and becomes more sophisticated over time (Battista, 2011; Clements & Sarama, 2007; Confrey et al., 2014; Lehrer & Schauble, 2015). An example of a progression within the measurement domain, which maps students thinking from a naïve conception of area and volume measurement towards an abstract understanding of the same, is Battista's (2004) learning trajectory (Figure 4). Battista (2004) has developed his measurement LT based on a series of preliminary studies. One study employed teaching experiments and interviews to establish the strategies that 3rd graders and 5th graders used when working on tasks that involve enumerating the cubes in 3D arrays (Battista & Clements, 1996). Later studies based on students' strategies compiled in the 1996 paper, further

developed the research on how students enumerate cubes in 3D arrays (Battista, 1999) and 2D squares (Battista et al. 1998).

Cognitive level	Description
Level 1	Absence of units-locating and organizing-by-composites processes
Level 2	Beginning use of the units-locating and the organizing-by-composites processes
Level 3	Units-locating process becomes sufficiently coordinated to recognize and eliminate double-counting errors
Level 4	Use of organizing-by-composites process to structure an array with maximal composites, but insufficient coordination for iteration
Level 5	Use of units-locating process sufficient to correctly locate all units, but less-than-maximal composites employed
Level 6	Complete development and coordination of both the units-locating and the organizing-by-composites processes.
Level 7	Students' spatial structuring and enumeration schemes become sufficiently abstract

*Figure 4.* Battista's (2004) LT taken from Lobato and Walters (2017, p.77).

One can see from Battista's (2004) trajectory that it is not simply a rational analysis of the domain as it provides descriptions of how students operating at different levels of thinking enumerate the number of cubes in an a 2D/3D array. For example, students functioning at Level 1 have yet to consider a row-by-column organization of units/squares and often count squares along a random path and consequently miscount. While students functioning at Level 5 have correct, elaborate strategies to enumerate the number of cubes but have not employed a counting strategy that is easily replicated in other cases or, in other words, is not a generalizable strategy.

Intermediate levels of an LT map a variety of epistemological objects such as partial conceptions, limited representations, common misconceptions, cases, properties, and strategies (Confrey and Maloney, 2017). However, the process of defining clear and useful intermediate descriptions between the bottom and top anchors of an LT, referred to as the “messy middle”

(Confrey, Maloney, Wilson & Nguyen, 2010) requires *extensive spadework*, in the words of Lehrer et al. (2014). The spadework typically entails detailed research syntheses and expert panel analyses which provide a basis for multiple phases of clinical interviews, teaching experiments, and/or design studies (e.g., Confrey et al., 2014; Clements & Sarama, 2004; Battista, 2004). Other researchers who focus on domains that already have an existing body of empirical research on student learning tend to base their preliminary LT on a review of extant literature (e.g., Arielli-Atalli, Wylie & Bauer, 2012; Carney and Smith, under review; Lai, Kobrin, DiCerbo & Holland, 2017; Pham, Bauer, Wylie & Wells, under review; Wilmot, Schoenfeld, Wilson, Champney & Zahner, 2011).

The methodology used to develop LTs has an influence on the grain size and level of clarity across levels within LTs. For example, an extensive research synthesis and a phase approach to clinical interviews afforded Confrey et al. (2014) the ability to describe how student conceptions and strategies progress along a 16-level LT for equipartitioning across 13 cases. Others such as Clements et al. (2014) provide a detailed developmental progression for early measurement topics alongside descriptions of student thinking in relation to a specific sequence of tasks. Lehrer et al. (2014) describe varying knowledge states, or cognitive landmarks, for 6 data modeling constructs that map LTs spanning between four and seven levels. This is not to say that level of clarity within the progress levels of LTs can only be achieved through multi-phase research projects.

Ultimately, researchers' vision for how they intend their LT(s) to be utilized by practitioners has an impact on the degree to which progress levels of LTs are articulated. LTs that are designed to inform day-to-day instruction (e.g., Battista, 2004; Clements and Sarama, 2004; 2007; Confrey et al., 2014; Ellis et al., 2016, Lehrer et al., 2014) may tend to provide more

detailed descriptions of student thinking in relation to a learning target than LTs which inform instruction on a less frequent basis. An example of the latter, is Wilmot et al.'s (2011) functions learning progression (Figure 5) which on its own does not articulate specific student strategies in reference to specific mathematical functions. Instead they adapt the SOLO taxonomy language to provide teachers with broad descriptions of students' responses to items on an assessment for college readiness e.g. *Level 3 is defined as (Multistructural) makes connections across multiple representations.*

Functions Learning Progression: Making Connections Across Multiple Representations of Mathematical Functions		
Levels of Sophistication:	Student is able to:	Response to items
<b>Extended Abstract (5)</b>	Make connections not only within the given subject area, but also beyond it, able to generalize and transfer the principles and ideas underlying the specific instance.	<ul style="list-style-type: none"> <li>Predict, explain, and synthesize their understanding to a real-world context.</li> <li>Solve non-routine problems including non-algorithmic functions.</li> </ul>
<b>Relational (4)</b>	Demonstrate understanding of the significance of the parts in relation to the whole.	<ul style="list-style-type: none"> <li>Compare/contrast information given in multiple representations of functions to demonstrate understanding of the content.</li> <li>Recognize which representations to choose for the context.</li> <li>Select representations and move fluently among them to achieve a solution.</li> </ul>
<b>Multistructural (3)</b>	Make a number of connections, but the meta-connections between them are missed, as is their significance for the whole.	<ul style="list-style-type: none"> <li>Make connections across more than two representations (e.g. symbolic, tabular, graphic and verbal representations).</li> <li>Recognize more than one relevant feature of a functional relationship.</li> </ul>
<b>Unistructural (2)</b>	Make simple and obvious connections, but their significance is not grasped	<ul style="list-style-type: none"> <li>Make pair-wise connections between representations.</li> </ul>
<b>Prestructural (1)</b>	Acquire bits of unconnected information, which have no organization and make no sense.	<ul style="list-style-type: none"> <li>Interpret graphs where both variables have to be interpreted, or where time as independent variable.</li> <li>Demonstrate understanding of function in one representation.</li> </ul>
<b>Prealgebraic (0)</b>	Acquire prerequisite skills	<p>Demonstrate understanding of:</p> <ul style="list-style-type: none"> <li>functional dependence, where a change in one variable affects another variable.</li> <li>continuous variables (time, distance) and dichotomous variables (hot or cold).</li> <li>Relative order and measurement to define variables.</li> </ul>

*Figure 5.* Functions Learning Progression taken from Wilmot et al. (2011, p. 265).

Researchers' models of how they expect students to "climb" up an LT also influence the level of detail that they provide when articulating their LT. For example, Clements and Sarama (2004) envisage that their LTs in K-2 mathematics education to be used along with a LT-based

curriculum (Clements & Sarama, 2007). According to Clements and Sarama, progress levels delineated in their LTs are tied to a specific sequence of tasks in their curriculum. They also believe that the learning process is somewhat incremental and that a student must construct a critical mass of the ideas at one level before “the thinking characteristic of the subsequent level becomes ascendant in the child’s thinking and behavior” (Sarama & Clements, 2009, p.21). Lehrer et al. (2014), on the other hand, have not prescribed a sequence of tasks but have developed curricular units based on their *Data Modeling* LTs to support the professional development of teachers as they learn to incorporate the LTs in their teaching repertoires. Lehrer et al. (2014) believe there are task-dependent conceptual pathways that students take in order reach a sophisticated level of understanding of statistical concepts. Battista’s (2011) position on student thinking in relation to the tasks employed during instruction is similar, but he believes that learners may not show reasoning as hypothesized by the entry point of the LT. That is, learners may enter at different levels of the LT and that their transitions may not be level-by-level or indeed always advancing. Confrey (2015) does not prescribe a sequence of tasks or provide teachers with curriculum to be used in support of her LTs but, like others, does believe that tasks play an important role in mediating students’ understanding of a mathematical concept. In her previous work with turnonccmath (2018), Confrey supported her LTs with illustrative tasks for progress levels and encouraged teachers to experiment with a variety of curricular approaches (Learning Trajectories based Instruction (LTBI) or project-based) and that one can use LTs to inform these various designs. And so, for Confrey using LTs to inform assessment is another possible option for supporting instructional planning. According to Confrey et al. (2017) data from LT-aligned assessments will help teachers diagnose gaps in students’ learning but also help them ascertain if their choice of curriculum is learner-centered and meeting the needs of

their students. Other researchers also recognize the importance of helping teachers drive instructional decisions using appropriate data sources and view assessments tied to LTs as sources of such data (e.g., Battista, 2004; Ellis et al. 2016; Wilmot et al, 2011) but most focus on single LTs and only few (Nichols et al, 2016) approach a system-based approach as Confrey does.

Confrey and her team were interested in how to get information about LTs to affect practice at scale. In order to do this, she built on her prior work with turnonccmath. With the turnonccmath, she had displayed the CCSS-M as a set of LTs but in her new web-based application called *MM 6-8* LTs can be associated to one or more of the CCSS-M. *MM6-8* is designed to: link to any form of curricular intervention, be a tool that is a learning map visualization of the content of middle school mathematics, and assess its outcomes using a diagnostic assessment. The LTs underpin all features of *MM 6-8*. Because this study is set in the context of this web-based application, it is necessary to review the underlying research premises of it before providing more detail on the focus on this thesis, which is on the validation of one particular LT in the map.

### ***Math-Mapper 6-8: A middle grades digital learning system***

With the growing prevalence of digital technologies, it is no surprise that new models of curricular development and delivery, professional development, pedagogy, and assessment are emerging (Dede, 2014). One form of these innovations, briefly described in the previous section, is a digital learning system (DLS) called *Math-Mapper 6-8* (*MM 6-8*). Confrey et al. (2017) defined a DLS as “a single platform into which learning targets, learning opportunities provided by digital curriculum materials and tools, assessment activities, and analytic capacity are all situated” p. 719. Her aim was to address the growing concerns in mathematics education of a

lack of curricular coherence (Larson, 2016) brought about due to teachers' increasing reliance on web-based curricular materials to address the new, CCSS-M (Confrey et al., 2017). One of the main concerns with teachers' use of web-based resources was that research had shown that teachers were not focusing on important cognitive intentions of the resources but instead focusing on secondary matters such the enjoyability of a task (Webel, Krupa & McManus, 2015).

Thus, *MM6-8* has been described as having the following six purposes: 1) to inform teachers of class progress by LT, 2) to elicit and leverage diverse levels of student thinking, 3) to increase students' awareness of their own learning and growth, 4) to strengthen teachers' content and pedagogical knowledge, 5) to connect instructionally proximal data to more distal forms of assessment (interim and high stakes), and 6) to improve student's knowledge of target constructs and big ideas from the map Confrey & Toutkoushian (in review, p.9). The web-based application is comprised of: 1) a learning map that describes the terrain of middle school mathematics to be learned; 2) a set of curated, open-source curricular materials, 3) a diagnostic assessment system, and 4) an underlying analytics system that provides researchers with metadata about the usage of the DLS.

To achieve these purposes, *MM 6-8* integrates several elements of instruction and leverages the responsiveness of digital technologies to facilitate data-driven decision making. *MM 6-8*'s assessment and reporting system has LT-aligned assessments which are digitally administered and scored in real time so as to provide both teachers and students with timely and actionable feedback. Assessment results are designed to be used formatively, so that both teachers and students have access to information about her/his class' performance on the LT and

what they need to do in order to improve their understanding of a concept (Brookhart, 2016; Heritage, 2007; Heritage, 2010).

LTs are what distinguish *MM 6-8*'s diagnostic assessments and the nature of feedback they provide to teachers and students from other classroom assessments (Confrey & Maloney, 2012). Feedback from typical assessments that are aligned to standards such as the CCSS-M report on achievement, i.e. whether a student understands a standard or not (Corcoran et al., 2009), whereas LT-aligned assessments enable teachers to diagnose students' understanding on a scale of increasing sophistication. Since the feedback from *MM 6-8*'s assessments are tied to the progress levels of the associated LTs, a teacher can "respond appropriately to evidence of their students' differing stages of progress by adapting their instruction to what each student needs in order to stay on track and make progress toward the ultimate learning goals" (Corcoran et al., 2009, p.19).

To this end, the suggested use of *MM 6-8*'s assessment system is outlined by the following scenario. When a teacher teaches a concept or a topic in mathematics, they are cognizant of the reality that some students come to class with varying conceptions about the target concept (e.g. partial conceptions, misconceptions, systematic errors (Confrey, 1990)) and therefore may have varying degrees of understanding of this concept(s). In order to help students take stock before they move on to another unit of instruction, the teacher administers an LT-aligned diagnostic assessment. Students take the diagnostic assessment knowing that the results will generate information about what they need to focus on in order to gain proficiency in that topic, rather than an evaluative judgement of whether they can or cannot "do" a certain topic. The teacher then receives whole-class data for each construct/LT in the relational learning cluster (RLC), which allows her/him to make inferences about individual students, subgroups of

students, or the class as a whole. The teacher report in *MM 6-8* includes a class-wide “heatmap”, so that in addition to displaying data about individual students, it also informs teachers about what progress levels to reteach to the whole class, how to form subgroups to meet specific needs, and how to decide when students are ready to proceed (Appendix C). Students also receive a report immediately after completing an assessment. The report consists of overall percent correct feedback on the assessment and item-by-item feedback presented in an “item matrix” (Appendix D). Students are also able to view their responses to items in relation to the progress levels of the LT. This additional feedback goes beyond a simple percent correct and should inform students about their particular strengths and weaknesses in relation to the LT progress levels. Students can then use this detailed feedback as a recommendation of where they should focus their efforts next in order to improve. *MM 6-8*’s practice feature is integrated with the diagnostic assessment system and provides students with recommendations for where, i.e. at what levels in the LT, they need to focus on during their practice session.

*MM 6-8*’s diagnostic assessments, embedded within the learning map, are designed to provide timely, actionable, and user-friendly feedback at the level of an RLC, so that 1) teachers can make informed decisions about instruction to support the learning of all students, and 2) students will better understand what they have learned, what they know, and what they do not yet know. Critical to the usefulness of such assessments is the idea that even though they are markedly different from (but mutually supportive of) large-scale, summative assessments (Shepard, 2006), they allow users to become systematic in their consumption of data because users have the assurance that:

- a) classroom assessments are designed to be relevant i.e. the assessment is comprised of questions that can measure student understanding alongside a continuum of learning

- towards a specific learning target (and within the framework of state/national educational standards).
- b) data generated about students along the same continuum of learning helps both teachers and students to reliably identify gaps in learning. Thus, the data informs their next steps as they work together to negotiate a path to close these gaps.
  - c) data generated from classroom assessments is transparent so that students can become partners in the learning process.
  - d) creating classroom norms around the ongoing use of a feedback loop prepares students for long-term learning goals (e.g. coverage of grade-level standards) and helps them learn about their own learning (meta-cognition).

To this end, developers of LT-aligned assessments need to offer both/all end-users evidence-based assurance that data from these LT-aligned assessments can indeed meet the users' needs as stated above. Evidence-based assurance, according to Heubert & Hauser, (1999) can come from establishing that an assessment is reliable, valid, and fair. A review of the broader assessment validity literature is examined in the next section so as to situate the validity efforts of researchers working on validating LT-aligned assessments.

### **Assessment Validity**

It is important to consider what LTs have to offer instruction as well as their implications for mathematics assessments, which function in close proximity to instruction (Shepard, 2006), to address calls for improving educational outcomes through systematic, formative assessment practices (Brookhart, 2016; Heritage, 2008; Black & Wiliam, 1998). At present, traditional summative or high stakes assessments, albeit necessary bookends of an accountability system (Confrey & Maloney, 2012), are failing to deliver the kind of data that teachers need to plan their

daily or weekly instructional sequences or units (Mislevy, Steinberg & Almond, 2003). Conversely, data from classroom-based assessments that accompany LTs can be used by teachers to diagnose any gaps in students' understanding on an ongoing basis, provide students with feedback in relation to a well-articulated goal, and involve students in monitoring their progress along a continuum of learning (Confrey & Maloney, 2012; Heritage, 2007; 2008). Conceptualizations of validity in the context of large-scale assessments should be considered first to see how they have influenced the dialogue on validating classroom assessments, especially those aligned to LTs.

Common types of assessment validity include *content*, *criterion-related*, *construct*, and *consequential* validity (Messick, 1989). *Content* validity is typically evaluated by expert analyses that check for alignment between the measurement instrument and the target domain. In the case of an LT-aligned assessment, a student's performance on a sample of items from the LT domain is evaluated, and the analyses judge whether the observed performance is an appropriate estimate of overall performance in the domain (Guion, 1977). *Criterion-related* validity examines relationships between an assessment score and other measures that should correlate to it (the criterion) (Messick, 1995). For LTs, for example, one can compare the outcomes of a locally defined, valid, end-of-unit test to the scores on the LT-aligned assessment. *Construct* validity seeks to establish that an assessment measures what it purports to measure (convergent validity) and that it does not measure irrelevant attributes (discriminant validity) (Messick, 1989). Regarding LT-aligned assessments, the question is whether an LT gives relevant, reliable, and accurate feedback concerning the student's standing within an LT. *Consequential* validity looks at the broader social consequences of using the outcomes of a particular test for a particular purpose (Downing & Haladyna, 2006). In the case of LTs, consequential validity can be

addressed by answering the question, *Does the nature of the feedback provided in the LT score reports improve teaching and teacher knowledge?* For Messick (1989) and other validity theorists, approaching the different validities in a fragmented manner without an underlying framework permits researchers to “pick and choose” based on the evidence that they have available (Kane, 2013). This can lead to inaccurate or biased assessments. To counteract inaccuracies and bias, Messick (1989) weaves together the various validities to provide a unified view of validity, one that he defines as "...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment" (Messick, 1989, p. 13). His definition follows from Cronbach’s (1971) notion that validity is not a property of an assessment or an instrument, but rather of the belief that to validate an assessment is to validate the inferences made from the test scores, and how those scores are used in a particular context.

To facilitate the use of Messick’s unified model on a more practical basis, Kane (2004) developed an argument-based approach to validation in which an interpretive/use argument (IUA) is developed by the test-developer, one to be used as the framework for the validation process. Claims asserted in the IUA require warrants that are supported by evidence. IUAs thereby provide test developers with a scheme for collecting appropriate and relevant data to support their claims (Kane, 2013). In subsequent literature, Kane’s (2004) argument-based approach to validation has shown promise in terms of facilitating the use of Messick’s construct model in a practical but rigorous manner. For example, Pellegrino et al. (2016) build on the Kanian framing of a “reasoned argument backed evidence” (2004, p.3) as well as the multiple facets of validity to develop a framework that specifically focuses on the interpretive uses of classroom assessments.

Pellegrino et al. (2016) provide an organized, evidence-based approach that conceptualizes the validity of instructionally relevant assessments into three main components: cognitive validity, instructional validity, and inferential validity based on Pellegrino et al.'s 2001 *Assessment Triangle* (Figure 6). In Pellegrino et al.'s Conceptual and Evidentiary Framework, *cognitive* validity evokes Messick's idea of *construct* validity but emphasizes the importance of a "scientifically credible" theory of how learners' understanding of concepts grows in sophistication. *Instructional* validity of an assessment is concerned with how the assessment is aligned to the standards of teaching (*content*) and whether the assessment meets the needs of the end user i.e. teachers, students, and school administration (*consequential*). Lastly, inferential validity acts as a check on whether the assessment is reliable in generating model-based predictions about student understanding that are not contradictory to information produced by other formal assessments that the students typically takes (*criterion*).

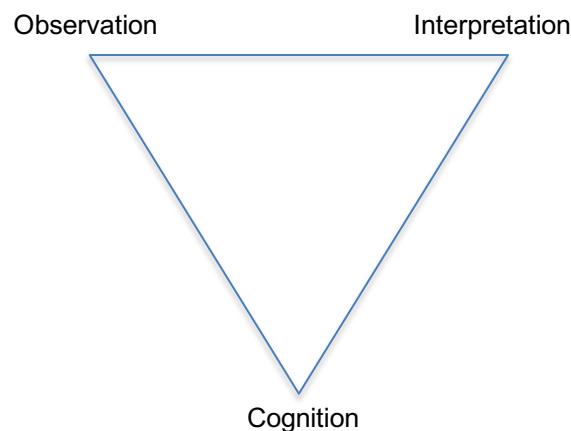


Figure 6. The assessment triangle taken from Pellegrino et al. (2001, p. 44).

### **Measuring and validating LT-aligned assessments.**

As highlighted within the review of the assessment validity literature, it is important for assessment developers to consider a variety of sources of evidence when validating assessment scores. The collection of different types of evidence; however, is not a haphazard occurrence.

Instead it is done in relation to a validity argument framework which outlines, a priori, the claims being made assessment developers and how they intend to support them. Such approaches to assessment design and validation are generally referred to as Principled Assessment Design and include within them a variety of methods such as Principled Design for Efficacy (PDE) (Nichols et al., 2016), Evidence Centered Design (ECD) (Mislevy et al., 2003) and the Berkeley Evaluation and Assessment Research (BEAR) method (Wilson & Sloane, 2000).

In mathematics education, many researchers follow the BEAR method when designing and developing LT (or LP) based assessments (Carney and Smith, in press; Clements, D.H., Sarama, J.H., & Liu, X.H., 2008; Confrey, Maloney, Nguyen & Rupp, 2014; Lehrer, R., Kim, M, Ayers, & Wilson, 2014, Wilmot, Schoenfeld, Wilson, Champney & Zahner, 2011). The studies that follow the BEAR method tend to report evidence that the items address the progress levels of the trajectory, and on forms of psychometric modeling (IRT Rasch models and Wright Maps) to demonstrate that the overall structure of the LT is modeled by moves from easier to harder items up the levels. Fit statistics are used to determine if items at the same level of the LT were behaving similarly (Szilagyi; Lehrer et al., 2014; Confrey et al., 2014). The BEAR method to validation is based on an item response modeling approach to constructing measures of learning progressions. Below are Wilson's (2005) four principles of the method, illustrated briefly though Lehrer et al.'s (2014) *Data Modeling* learning progression (LP) below (Figure 7).

<b>Construct</b>	<b>Initial Performance</b>	<b>Intermediate Performance</b>	<b>Highest Performance</b>
Data Display (DaD)	DaD1(a): Create or interpret displays without relating to the goals of the inquiry.	DaD4(a): Display data in ways that use its continuous scale (when appropriate) to see holes and clumps in data.	DaD6(a): Discuss how general pattern or trends are either exemplified or missed from a subset of case
Meta-Representational Competence (MRC)	MRC1(a): Recognize that displays represent data, but misinterpret one or more elements of the display.	MRC3(a): Compare displays by indicating what each shows about the structure of the data.	MRC5(b): Coordinate different display types and formats to strengthen an argument.
Conceptions of Statistics (CoS)	CoS1(a): Use visual qualities of the data to summarize the distribution.	CoS3(d): Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among components.	CoS4(d): Predict and justify changes in a sampling distribution based on changes in properties of a sample.
Chance (Cha)	Cha1(a): View chance as being under the control of an agent.	Cha4(c): Recognize that an “unlikely” string of outcomes is possible and even expected over many repetitions of the event (e.g., 3H run in 10 coin tosses).	Cha6(b): Use sample space to estimate probability and/or relative frequency of compound (aggregative) events.
Modeling Variability (MoV)	MoV1(a): Attribute variability to specific sources or causes.	MoV3(a): Use chance devise to represent variability.	MoV5(a): Judge model fit in light of variability across repeated simulation with the same model.
Informal Inference (InI)	InI1(a): Make a judgment or prediction according to personal experience or beliefs.	InI4(a): Make predictions based on regions of values such as clumps or holes, sometimes identified by particular values.	InI7(b): Compare two objects (conditions, etc.) by considering the variations in the sampling distributions of their statistics.

*Figure 7. Abbreviated constructs maps taken from Lehrer et al. (2014, p. 43).*

1. The assessment must be based on a developmental perspective i.e. student learning is measured over a period of time rather than a single point in time: the data modeling LPs synthesize research on student learning within the disciplinary practice of data modeling into six constructs (row headers) and associated learning performances/levels (column headers) which delineate how student learning progresses. Wilson (2005) refers to these as *Construct Maps*.
2. There must be a match between the instruction and the assessment: Lehrer et al. (2014) developed assessments items to meet each of the learning performance and also developed curriculum that supported the professional development of the teachers who

participated in their study. According to Wilson (2009) the framework for the assessments, curriculum, and instruction must all be the same.

3. Assessments must be manageable by teachers: Lehrer et al. (2014) field tested assessment items and then mined students' responses to these in order to develop scoring guidelines and exemplars of student work. Wilson (2005) refers to these as the *Outcome Space*. The scoring guidelines and student work were made available to teachers for use when administering the assessments.
4. There must be evidence of a high-quality assessment: Lehrer et al. (2014) analyzed data from field-testing using a psychometric (partial-credit, 1 dimensional item response (IRT)) model. The authors adopted an iterative design process and made a number of adjustments based on their analysis (e.g. developed new items when the original set of items did not fully cover a construct, revised or discarded items that did not appropriately measure the construct, and made some edits to one of the constructs and its levels).

Other researchers such as Confrey & Toutkoushian (in review) and Lai et al. (2016) subscribe to a class of PAD frameworks that apply Pellegrino et al.'s framework (2016). Confrey has adapted Pellegrino's framework to create a validation framework tailored to their DLS *MM6-8*. A summary of *MM 6-8*'s framework is given below to show how she structured her validity argument across the three components from Pellegrino et al.:

*Cognition:*

1. LTs are linked to significant and worthy targets, related to the big ideas and standards;
2. LTs exhibit clear and comprehensive coverage, describing commonly elicited or observed student behaviors;
3. Progress levels are ordered by increasing sophistication, with grain-sizes distinct enough to foster measurement;
4. Multiple entry points are supported;

5. LTs incorporate the probabilistic character of student progress;
6. LTs facilitate upwards movement via instruction.

*Instruction/Implementation:*

1. LT-based assessments are aligned coherently with curriculum scope and sequences, and with instruction;
2. Students and teachers conduct review of class and student reports;
3. Teachers orchestrate post-assessment follow-up, based on prior assessment data; and
4. Groups of teachers adjust and revise subsequent curricular approaches based on assessment outcomes.

*Inference:*

Analytic methods from measurement and statistical inference are applied to analyze data from students' task performances, to discern the degree to which they "reliably align with one or more underlying conceptual measurement models that are appropriate to the intended interpretive use" (*ibid.*, p.7-8 ).

Confrey and Toutkoushian (under review) conducted a validation study applying the aforementioned validation framework to one particular learning trajectory on *Measurement of Characteristics of Circles*. Their method distinguishes three categories of variability affecting the difficulty of items and use these in their analyses of IRT item-difficulty plots (p.11):

1. Variation in difficulty among items that reside in a single progress level within an LT: intra-level variation ("Intra-LV"); such items may address different facets of reasoning at a single level.
  2. Variation in item difficulty between levels within an LT: inter-level variation ("Inter-LV").
  3. Variation associated with bias and noise: construct-irrelevant variation ("Irrel-V"). Sources may include ambiguous wording, a poor or misleading representation, or students' lack of familiarity with a problem context.
- Reports of consequential and criterion-level validity are reported more infrequently.

Only a few studies report on in-depth qualitative analyses of items using "think-aloud" interview data as empirical evidence supporting the *substantive* aspect of construct validity, i.e., "evidence that theoretical processes are actually engaged by respondents in the assessment tasks" (Messick, 1995, p.745). For example, Wilmot et al. (2011) conduct think-aloud studies for

assessment items but in the context of a six-level learning progression that spans several grade levels. That is to say that the level of detail about student thinking afforded at each level of their LP is different to say, the level of granularity of the LTs in *MM 6-8* such as the *Similarity and Scaling* LT.

### **Setting up the Study**

As indicated in this review, there is a need for investigating the validity of LT-aligned assessments which are designed to function in close proximity to instruction. While, the research on validity in the context of large scale assessments is not completely practical, it can influence the conceptualizations of validation work that is being done in the context of classroom assessments. However, researchers developing these assessments will be required to pay heed to more than just the IRT item difficulty characteristics of items. Measurement theory alone cannot help to validate such assessments. While it is apparent that IRT can help researchers scrutinize data to identify items that do not conform to the structure of an LT, it is yet to be determined how this approach impacts the fate of items that measure the kind of flexibility in thinking required of mathematics students, or more importantly the integrity of the LT itself. The assessment items and the LT can be examined by integrating a learning science perspective in the analysis of quantitative data (Confrey & Toutkoushian, under review) or using qualitative data from think alouds as “collective evidence of student understanding” (Wilmot et al., 2011, p. 267).

The design for this empirical study was guided by the gaps in previous research. The aim of this study is to use both quantitative and qualitative approaches in the context of an LT-aligned assessment that is designed to be used by middle grades’ teachers and students to

identify gaps in students' knowledge about geometric similarity so that they can work together to close these gaps.

### **Chapter 3: Methodology**

This research study is situated within an NSF grant-funded project titled, *Building a Next Generation Diagnostic Assessment and Reporting System within a Learning Trajectory-Based Mathematics Learning Map for Grades 6-8*, carried out by the Scaling Up Digital Design Studies (SUDDS) group under the direction of Dr. Jere Confrey (DRL-1621254). The SUDDS team has been working to develop and study the implementation of a digital learning system (DLS), based on the theoretical framework of learning trajectories, called Math-Mapper 6-8 (MM 6-8). The DLS consists of a learning map, organized around big ideas in middle grades mathematics, a set of illustrative instructional resources, and a set of diagnostic assessments designed to measure students' progress along learning trajectories within 24 clusters (defined in the next section). One of the major research efforts of the group is to study the approaches to validating a new genre of diagnostic assessments which are: a) tied to a theory of learning, and b) designed to be used in close proximity to instruction i.e. on a formative basis.

The purpose of this study was to investigate the validity of one learning trajectory-based assessment developed for students in the middle grades who are studying geometric similarity. The validation work included analyses of data collected over a two-year period from participating schools, combined with an external assessment and think-aloud interviews. These multiple data collection tools are used to gather evidence to support four separate but connected validity claims about the assessment. The goal of the study is to illustrate how to support validity claims in the context of a single digitally-administered classroom assessment that is aligned to a

learning theory. Furthermore, in carrying out this approach to validation, another goal is to inform the field on how to conduct validation studies in the context of this genre of assessments.

This chapter begins with the research questions, their associated validity claims and a brief description of the DLS and its features. Next, the methods and sample population are described, followed by a description of the data sources and data analyses used to answer each of the research questions.

## **Research Questions**

**Overall question:** To what extent do the assessments aligned to the *Similarity and Scaling* learning trajectory within the digital learning system *MM 6-8* provide a user with valid information about student thinking along a continuum of learning? The following research questions are considered to address this overall question:

1. To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in *MM 6-8*?

**Claim:** The items in the diagnostic assessments that measure students' understanding of geometric similarity are aligned to the corresponding LT levels as defined in the DLS *MM 6-8*.

2. To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?

**Claim:** Within the target population, item difficulty increases as the LT-level increases and the item difficulty of multiple items at a given level vary due to instructionally appropriate, construct-relevant factors.

3. How do patterns of observed performance differ across students from a diverse sample on LT-based assessments measuring geometric similarity?

**Claim:** The diagnostic assessments that measure students' understanding of geometric similarity are equally sensitive to the construct-relevant student thinking of a diverse sample of students.

4. How do the students' results on the *MM 6-8 Similarity and Scaling* assessment compare to those on an independent assessment measuring geometric similarity?

**Claim:** Scores produced by diagnostic assessments that measure LTs are correlated to scores produced by other assessments that measure student understanding in the same domain.

## Materials and Tools

### Math-Mapper 6-8.

*MM 6-8* represents learning targets by way of a hierarchical learning map (Figure 8) of nine big ideas (Appendix E) rooted in four “fields” (quadrants). As users zoom “into” the map they can access greater levels of specificity (Figure 9).

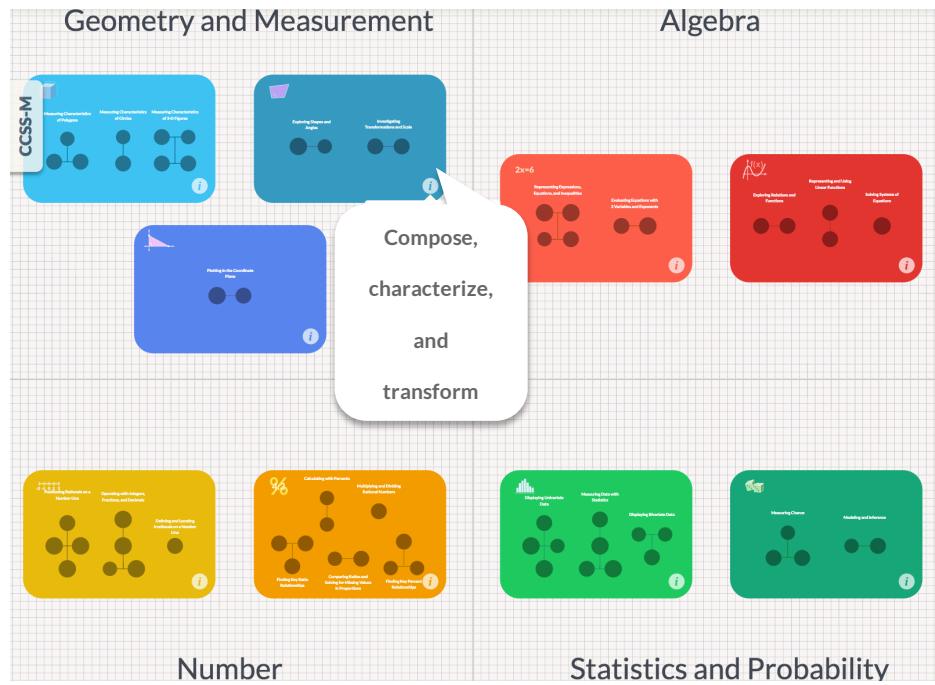


Figure 8. Math-Mapper’s learning map.

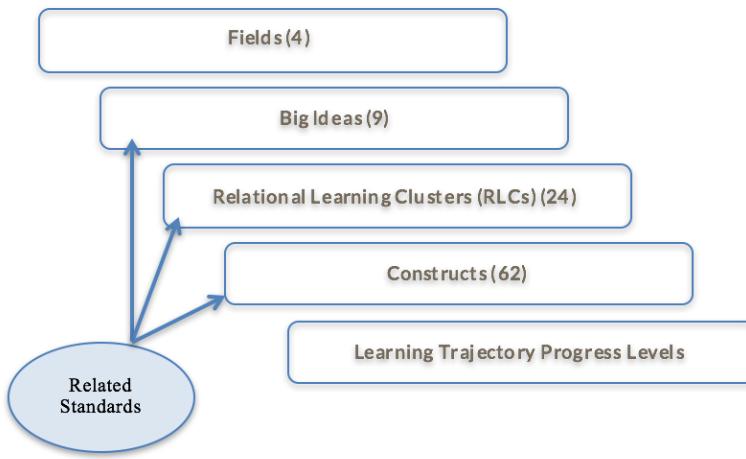


Figure 9. Hierarchy of elements that form Math-Mapper’s learning map.

For example, the big idea “*Compose, characterize, and transform lines, angles, and polygons*” shown in Figure 8 is a broad learning target that spans multiple grades. Within this big idea, users can zoom-in to see *clusters* of one to four related *constructs*: key concepts that should be learned in relation to each other. These clusters of constructs are called Relational Learning Clusters (RLC). In addition to informing teachers about how to group related concepts such that

students are able to have a coherent learning experience, RLCs also serve as the unit of testing in *MM 6-8*'s diagnostic assessment system. In the case of the RLC titled *Investigating Transformations and Scale*, the two constructs that should be learned in relation to each other are *Similarity and Scaling* and *Transformations of Geometric Shapes* (Figure 10). A user can click/tap on each construct to view the LT associated with it. The LT describes how students' observable behaviors increase in sophistication over time.

### **The *Similarity and Scaling* LT in *Math-Mapper 6-8*.**

The LT on geometric similarity proposed by Confrey (2015) in *MM6-8* was developed from existing empirical research (Chazan, 1988; Confrey, 1992; Cox, 2103; Lehrer, Strom & Confrey, 2002). The SUDDS team is unaware of any other LT on similarity. The preponderance of empirical work starting with very early work on student thinking in geometry to more recent work shows that students in pre-Kindergarten, elementary, and middle grades can come to quite a sophisticated appreciation of similarity given the appropriate opportunities to learn. Visual strategies that are promoted earlier in the LT support students as they progress towards proportional reasoning strategies.

Similarity and Scaling		
9	Uses informal arguments about angles to establish the AA criteria for similarity and solves similar triangle problems in context	8
8	Describes the effect of scaling a linear unit on area and volume	7
7	Uses similarity, proportionality, and scale factors to solve problems involving scale drawings	7
6	Extends similarity relationship from triangles to other shapes, and identifies similar shapes with proportional corresponding sides and congruent corresponding angles	7
5	Defines similar triangles as having equal angles and sides in proportion and, given one pair of corresponding sides, solves for unknown corresponding sides	7
4	Grows or shrinks a similar triangle by doubling, tripling, or halving side lengths and defines similar triangles as having congruent angles and proportional sides	7
3	Identifies similar triangles as having matching angles but possibly differing side lengths	7
2	Identifies similar objects as having the “same shape” and justifies by informally scaling, e.g. centering, or matching vertices	E
1	Identifies two shapes as exactly the same or congruent by placing one on top of the other	E



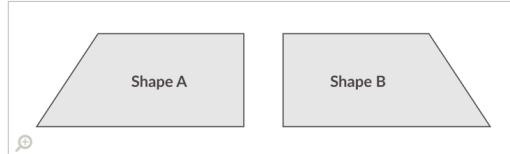
Transformations of Geometric Shapes		
5	Creates sequences of transformations to align figures, matching lines and angles	8
4	Connects dilations as a means to preserve similarity using a scale factor and can reverse	8
3	Uses angles to describe and quantify rotations, identifying the center of rotation as the vertex of the angle describing the rotation	8
2	Identifies a line of reflection and reflects figures over the line	8
1	Informally distinguishes among translations, reflections, and rotations	8

Figure 10. The two LTs associated with *Investigating Transformations and Scale*.

The LT begins by building similarity from this informal concept of congruence (L1) to a visualization of similarity as “the same shape but a different size”. For example, one of the items that assesses the first progress level<sup>1</sup> is shown in Figure 11. It assumes that students informally understand similarity and congruence by considering overlapping shapes (Confrey, 2015). However, it also takes into account a common student misconception at this level of the LT which is that two shapes must be oriented in the same way in order for them to be categorized as “the same” or congruent (Lehrer, Jenkins, & Osana, 1998). In the item presented in Figure 11, a student who responds by selecting Student 1 or 4 may believe that for Shape A and B to be congruent they must look exactly like each other. Student 5’s comment, on the other hand, is a compelling distractor for students who believe that equal areas are a sufficient condition for testing congruence of two or more shapes.

<sup>1</sup> Identifies two shapes as exactly the same or congruent by placing one on top of the other

Five students were discussing whether the two shapes below are congruent. Their comments are in the table below.



Student	Comment
Student 1	Shapes A and B are not congruent, because shape B is backwards.
Student 2	Shape A is congruent to shape B, because I can flip shape B and it fits exactly on top of shape A.
Student 3	Shapes A and B are congruent, because if I slide them together, they will fit into each other.
Student 4	Shapes A and B are not congruent, because they look different from each other.
Student 5	Shapes A and B are congruent, because they have the same area.

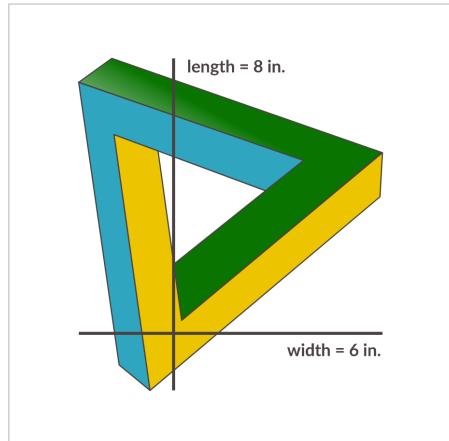
Which student is correct?

- Student 1
- Student 2 [1]
- Student 3
- Student 4
- Student 5
- None of these students are correct.

*Figure 11.* Assessment item aligned to Level 1 of the *Similarity and Scaling* LT.

Then, students engage in a variety of visualization actions as a means of comparing shapes by centering them (L2), or by examining how close they are to prototypical shapes (circular, equilateral, etc.). At the next level (L3), when working with triangles, students recognize that similar shapes have congruent angles, but not necessarily congruent side lengths. Students can then notice that they can build larger, similar versions of their triangles from their existing triangle and see that the larger triangles tend to have the same angles at the vertices but have side lengths that are  $n$  times as long (L4).

Diedre wants to put the logo below on the back of her cell phone. She measures the logo and finds that it is 8 inches long by 6 inches wide. Her cell phone is 6 inches long by 3 inches wide.



Diedre knows that she needs to shrink the logo for it to fit on her phone. Which size logo will fit on her phone and not be distorted?

- 2 in. long by 3 in. wide
- 5 in. long by 3 in. wide
- 4 in. long by 3 in. wide [1]
- 6 in. long by 4 in. wide
- 6 in. long by 4.5 in. wide

Figure 12. Assessment item aligned to Level 4 of the *Similarity and Scaling LT*.

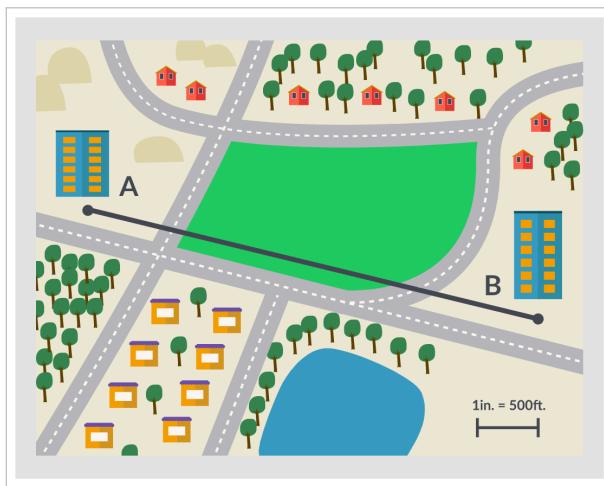
Figure 12 shows an item that is aligned to Level 4<sup>2</sup> of the LT. This item is written with the assumption that students recognize that in order to avoid distorting a shape when re-sizing it, the dimensions of the shape must be scaled in proportion. Students who select the second and the fourth answer options in this item may be using additive reasoning instead of multiplicative reasoning. On the other hand, students who select the last answer option may be using correct multiplicative reasoning but not within the context of the question.

At Level 5, students also recognize they can split the sides by  $n$  to create similar triangles and are then able to manipulate the sizes of triangles by using a combination of stretching and shrinking. In addition, as students come to understand that a fractional operator of  $\frac{a}{b}$  means  $a$  times as long and  $b$  times as short, then they can make a triangle with any given side length through proportional reasoning. At Level 6, they extend their understanding of similarity to other shapes, and use similarity, proportionality, and scale factors to solve problems involving scale drawings (L7).

<sup>2</sup> Level 4: Grows or shrinks a similar triangle by doubling, tripling, or halving side lengths and defines similar triangles as having congruent angles and proportional sides

Given that the *Similarity and Scaling* LT addresses 7th grade content, the assumption is (in the context of *MM 6-8*) that students will have studied Key Ratio Relationships and will be able to compare ratios and thus find missing values in proportions (6th grade content). If students have had the opportunity to learn the 6th grade content, then we would expect students to use proportional reasoning to solve Level 7 items. For example, for item 770 (Figure 13), we may see them using the following strategy: they may verbalize that 1 inch represents 500 feet is a ratio. Then, we expect them to set up a ratio box (Figure 14) and recognize that they can either multiply 500 by 7.2 OR multiply 7.2 by 500 to find their answer. They may also recognize scales of the form 1 in represents 500 feet as unit ratios and thus, use the unit ratio to build up or down to a given value.

A map is drawn with a scale as shown, where 1 inch represents 500 feet. Buildings A and B are 7.2 inches apart on the map.



Find, in feet, the actual distance between the two buildings.

[1] [3600.0 ≤ x ≤ 3600.0]

Enter a number

Figure 13. Assessment item (item 770) aligned to Level 7 of the *Similarity and Scaling* LT.

Scale drawing (inches)	Real life (feet)
1	500
7.2	?

Figure 14. Example of a ratio box representation used for solving item 770.

Students who have covered 7<sup>th</sup> grade proportional reasoning content<sup>3</sup> prior to the *Similarity and Scaling* LT may use a different approach to solve the same problem. For example, students who have had the opportunity to represent proportional relationships between two quantities using graphs ( $y = kx$ ) may recognize that  $k$ , the constant of proportionality can be adopted as the scale factor when working in the context of scale diagrams. Students should also be able to explain what any given points on the graph represents. If the scale given in a drawing is of the form  $1 : n$ , students should be able to explain that “1 unit of length on the scale diagram represents  $n$  units of length in the real world.” If the scale given in a drawing is of the form  $n : m$ , then students should be able to find one of the two units ratios ( $1 : \frac{m}{n}$  or  $\frac{n}{m} : 1$ ) and then use a unit ratio to find a proportional relationship between the scaled and real lengths.

At Level 8, students recognize the area or volume of a scaled object is not a product of just the scale factor and the original area or volume. Finally, at Level 9 they draw on their knowledge of the angle sum of triangles to recognize that if a pair of angles of two triangles are congruent, then they must be similar, and their sides should be in proportion. An example of an item at Level 9<sup>4</sup> of the LT (Figure 15) is based in the assumption that students have been given

---

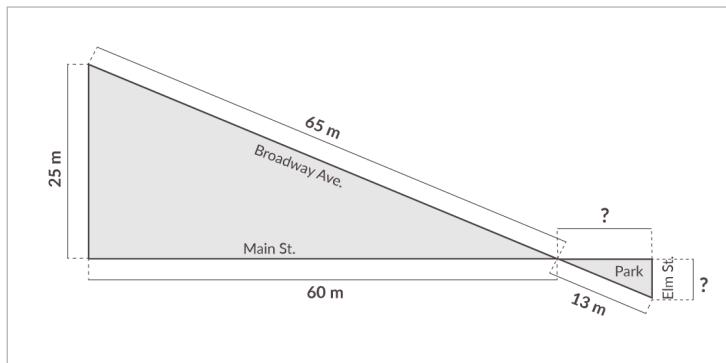
<sup>3</sup> Recognize and represent proportional relationships between quantities. Decide whether two quantities are in a proportional relationship, e.g., by graphing. Identify constant of proportionality from graph. Explain what a point on the graph of a proportional relationship means in terms of the situation. (7.RP.2)

<sup>4</sup> Level 9: Uses informal arguments about angles to establish the AA criteria for similarity and solves similar triangle problems in context

the opportunity to generalize the concepts described by levels 3<sup>5</sup> and 4 and are able to extend their understanding to solving problems with missing side lengths in a context.

In total, there are 21 field-tested items aligned to this LT. Details on how many items are aligned to each level of the LT, the types of items, and the number of responses observed per item are provided under the *data sources* section of this chapter. See Appendix F for the complete set of assessment items.

An architect is designing a park for a small triangular piece of land formed by the intersection of Broadway Avenue and Main Street. She needs the dimensions of the piece of land. She knows the dimensions of the larger triangle, as shown, and she knows the length of the park along Broadway Avenue is 13 meters.



What is the length of the park along Main Street?

[1]  $12.0 \leq x \leq 12.0$

Enter a number

What is the length of the park along Elm Street?

[1]  $5.0 \leq x \leq 5.0$

Enter a number

Figure 15. Assessment item aligned to Level 9 of the *Similarity and Scaling* LT.

## Method

Addressing the overall research question for this study (*To what extent do the assessments aligned to the Similarity and Scaling learning trajectory within the digital learning system MM 6-8 provide a user with valid information about student thinking along a continuum of learning?*) not only requires multiple sources of data but requires the researcher to integrate the interpretation of the results from each of four research questions as explained below.

Assessment items aligned to the *Similarity and Scaling* LT are digitally administered at field testing sites via the DLS MM 6-8. Data generated from these LT assessments can only be

---

<sup>5</sup> Level 3: Identifies similar triangles as having matching angles but possibly differing side lengths

analyzed using psychometric models such as Item Response Theory (IRT). The quantitative analyses can provide empirical evidence for one of the validity claims (*Research question 2*) i.e. item difficulty does (or does not) increase monotonically across the progress levels in an LT. However, as Confrey and Toutkoushian (under review) note, difficulty of items at a given level may vary due to mathematical cases. In the context of geometric similarity, students' performance may vary when solving problems that have whole number scale factors versus those that have fractional scale factors, even if fractional scale factors are encountered in lower level items. In addition, students' performance on items is also variable based on the opportunities they have had to learn the content. That is to say, variability in item difficulty between items at a level (or lack of variability between items across levels) is not always an indication of the failure of an item to measure the LT level appropriately. To account for variability between items at a level and variability across levels, in a systematic manner, qualitative methods such as data from think-aloud interviews with students are needed to interpret findings from the quantitative analyses. A mixed methods design allows the researcher to better understand how a representative sample of students think about and respond to assessment items aligned to the *Similarity and Scaling* LT from both a learning science perspective and a measurement (psychometric) perspective.

Data collected to answer the first research question pertaining to students' responses to assessment items during think-aloud interviews, were analyzed using verbal protocol analysis (Ericsson & Smith, 1993). Qualitative results from the first research question then assisted in the interpretation of the findings from the quantitative analysis applied to the field test data. Data collected to answer the third and fourth research questions are analyzed exclusively using quantitative methods; analysis of the distribution of students' scores across ethnic subgroups and

across both genders, and the correlation of scores generated from an independent assessment with those generated from the *Similarity and Scaling* LT assessment.

Mixed methods research is conducted using a variety of approaches (Small, 2011). Research methods vary because scholars may opt to use qualitative methods to confirm or complement quantitative data, or vice-versa. Other trends in research indicate that scholars choose to collect different types of data simultaneously or sequentially (Johnson, Onwuegbuzie, & Turner, 2007; Small, 2011). This study specifically follows a mixed method convergent (concurrent) design model (Figure 16), which consists of multiple phases of quantitative and qualitative data collection (Creswell & Plano Clark, 2017). The data are analyzed separately but integrated at the interpretation and reporting level (Fester et al. 2013).

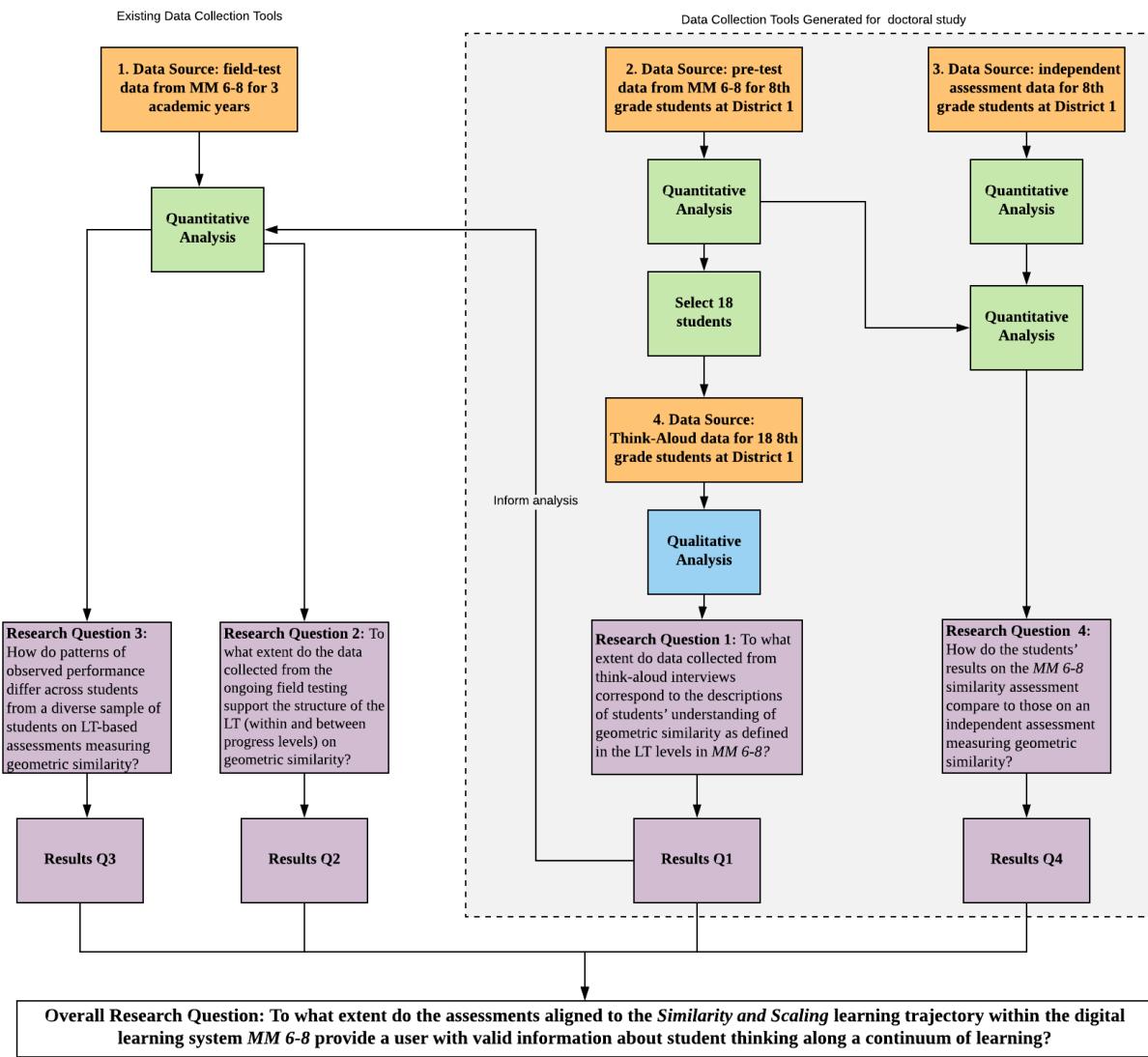


Figure 16. Convergent design: integration at the interpretation and reporting level.

## Sample

Data collection for this study came from three groups of students (Figure 17). The first, largest group of students are all the students participating in the field-testing of *MM 6-8*. These students come from three middle schools across two districts that have had an ongoing research partnership with the SUDDS group and have been field testing sites for *MM 6-8* since the 2015-16 academic year. Assessment data for two academic years from these schools has been used to answer the second and third research questions. Students who respond to items aligned to the *Similarity and Scaling* LT are typically enrolled in 7th grade mathematics courses but two of our partner schools offer an accelerated track for 6th grade students. Therefore, it is possible that some of the students who responded to these items are also from 6th grade. In either case, because teachers are responsible for managing the assessments, it is assumed that all students taking these assessments have had the opportunity to learn the content.

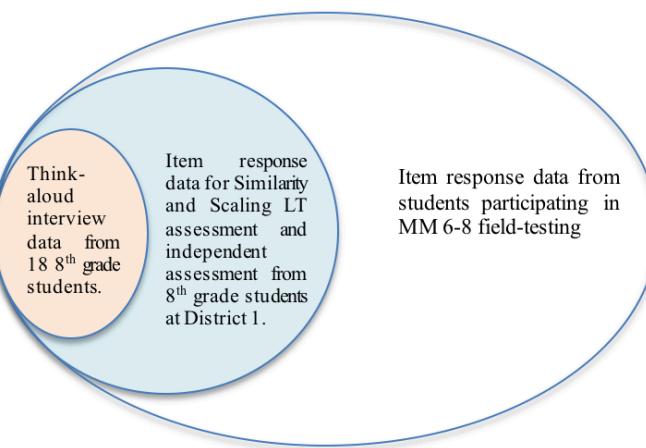


Figure 17. Groups of students who participated in the study.

Table 1 shows both districts' middle school demographic data. The school in District 1 is in its third year as a 6-8 school, is growing rapidly in its 1-to-1 technology use, and serves a

diverse, academically high-needs, student body. District 2 is wealthier, is high-performing, and has a longer history with student use of computers.

Table 1. Demographics for All Sites.

	District 1	District 2
Population served	977	1163
African-American (%)	27	4
Asian (%)	1	9
Hispanic and Mixed (%)	10	8
White (%)	53	79
Other (%)	10	-
Percent Free and Reduced Lunch	56.9	9.9
Number of years implementing 1-1 computing	3	5
Number of teachers participating	19	33
Number of tests taken	9,197	21,696

Subsets of the overall group of students from partner schools are used to answer the first and fourth research questions. In particular these are the 8<sup>th</sup>-grade students from District 1; 136 of them took an assessment aligned to the *Similarity and Scaling* LT. Following quantitative analyses of their assessment results, a purposeful sample of 18 students (Table 2) were selected to participate in the think-aloud interviews. The criteria for selecting these 18 students is explained further in the *data sources and analysis* section.

Students from the 8<sup>th</sup> grade classes at the middle school in District 1 were recruited because the student body is ethnically more diverse than the student body at District 2. Collecting think-aloud interview data from an ethnically diverse group of students would allow for interpretations of the data to be more representative of typical middle schools in the United

States. In addition, scheduling was a pertinent reason for selecting 8<sup>th</sup>-grade students from District 1: the 8<sup>th</sup>-grade teachers were scheduled to teach geometry in the second quarter of the year (mid-late November). Due to this schedule, it was possible to conduct think-aloud interviews with 8<sup>th</sup>-grade students earlier in November and make a fair assumption that they had opportunity to learn the geometric similarity content covered by the LT assessment being investigated for this study but had not begun studying geometric transformations. The study of geometric transformations is closely tied with the study of similarity and scale (Seago, Jacobs & Driscoll, 2010; Usiskin, 1972). Thus, the researcher wanted to ensure that the students being interviewed were demonstrating their understanding of similarity in the context of only the 7<sup>th</sup> grade content.

Table 2. Demographics for Think-Aloud Interviewees.

Ethnicity	Number of Males	Number of Females	Total Number	% of sample (School %)
African-American	1	4	5	28 (27)
Asian	1	0	1	<1(1)
Hispanic	1	2	3	16*
White	0	7	7	42 (52)
Mixed	1	1	2	1*

\*Hispanic and Mixed-race students are classified under one sub-group (10%) for the school-level data

## Data Sources and Data Analysis Procedures

Earlier in the chapter, Figure 14 showed the chronology of the data collection process and linked the various sources of data to the research questions. This section of the chapter is to detail each of the components that made up Figure 14. To begin with a table that maps out the data sources associated with each research question is provided below (Table 3). Then, the remainder of the chapter is organized in the following manner: first, each question is re-stated,

followed by the data sources that are used to answer the question, and a description of how the analyses were carried out.

Table 3. Mapping of Data Sources to Research Questions.

Research Question	Data Source
1. To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in MM 6-8?	<i>Think-aloud interviews</i>
2. To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?	<i>Field-test data from MM 6-8, Analyses of Think-alouds from Q1</i>
3. How do patterns of observed performance differ across students from a diverse sample of students on LT-based assessments measuring geometric similarity?	<i>Field-test data from MM 6-8</i>
4. How do the students' results on the MM 6-8 similarity assessment compare to those on an independent assessment measuring geometric similarity?	<i>Similarity and Scaling LT assessment, Independent Assessment</i>

### **Research question 1.**

*To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in MM 6-8?*

Responding to this research question allowed the researcher to determine if the assessment items aligned to the levels of the *Similarity and Scaling* LT were measuring student thinking as elaborated in the description of the LT levels. In addition, data from interviews allowed the researcher to establish whether there were patterns in student thinking that were not as described in the LT levels but still mathematically relevant. Furthermore, results addressing

this research question were integrated with results from field test data to answer *Research question 2*.

**Data Source: Think-aloud interviews.** Think-aloud interviews were conducted using an interview protocol (Appendix G) informed by Ericsson and Simon's (1993) protocol for using verbal reports as data. Think-alouds are also referred to as cognitive labs in the assessment literature (Zucker, Sassman, & Case, 2004) and are typically conducted in two major formats: concurrent verbalizations and retrospective verbalizations. Concurrent verbalizations require the participants to verbalize and not describe or explain what they are doing as they attempt a question. According to Ericsson and Simon (1993), social verbalizations such as explaining one's strategies for solving a mathematical problem demand different cognitive processes, therefore they recommend that social interaction be minimized during the think-aloud interviews. With this in mind, the researcher did not interrupt students as they verbalized their thoughts whilst attempting to answer an item except when they fell silent. If the student fell silent, the researcher reminded them to "continue thinking out loud." The verbal data collected have a closer correspondence to the actual processes students used to respond to the assessment question by minimizing social verbalizations; however, concurrent verbalizations have limitations in that data collected may not be coherent nor complete. Branch (2000) states that data from concurrent verbalizations are less useful when a task involves a high cognitive load or when the information is difficult to verbalize because of its form (i.e., visual data). Retrospective verbalizations are more complete than concurrent verbalizations and are conducted immediately after the task has been completed. The idea underlying retrospective verbalizations is that a portion of a train of thoughts which occurs as the student is responding to an assessment item is stored in the student's long-term memory. Immediately after the assessment item has been

answered the student can access this train of thought (Ericsson, 2006; Ericsson & Simon, 1993).

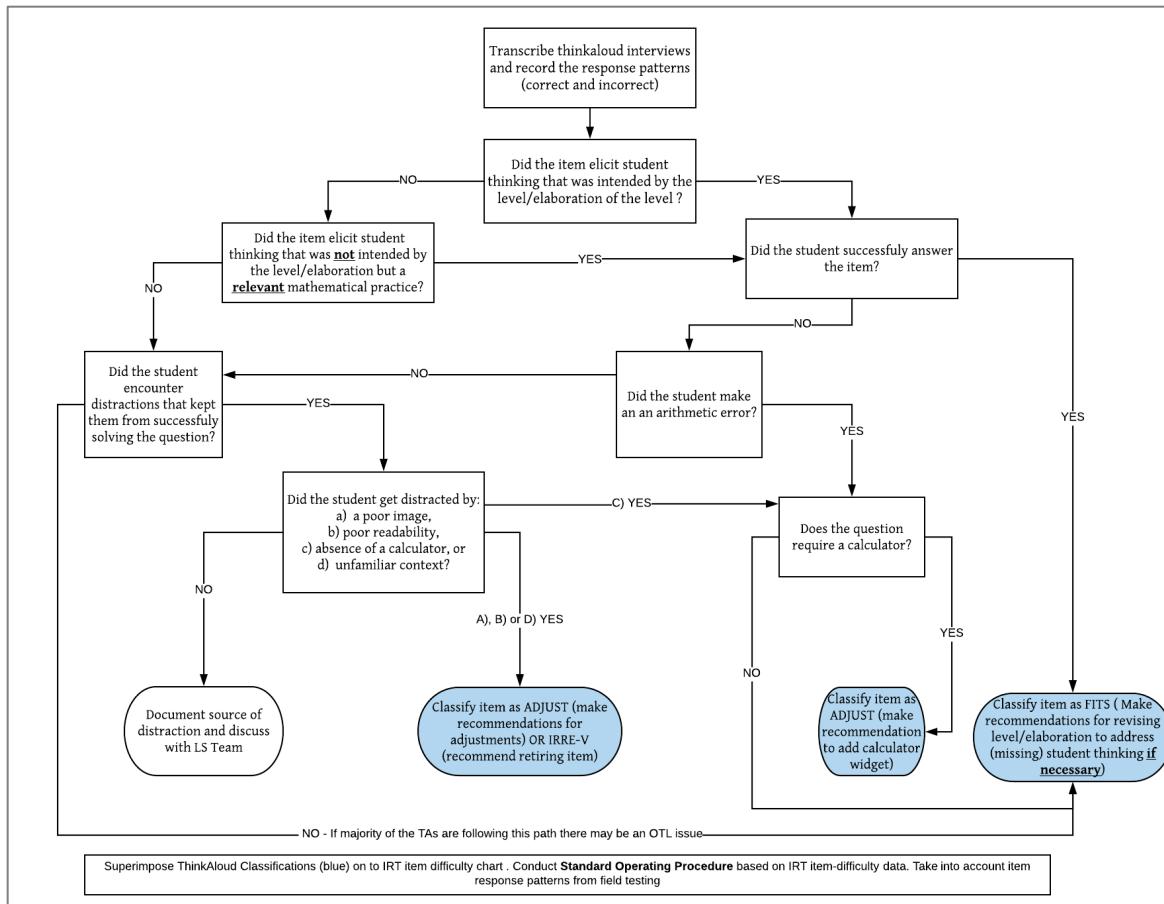
It was important for the researcher to establish whether the item was functioning the way it was designed to and document any productive mathematical strategies that students were using which were not already described by the LT level because a well-established LT for geometric similarity does not exist in the literature. For these reasons, both concurrent and retrospective verbalizations were employed for the think-aloud interviews conducted for this study. Thus, during think-aloud interviews social interactions between the researcher and student were limited while students worked the given items (i.e. concurrent verbalization), and then the researcher prompted students to reflect on their thinking while working once each item was completed (i.e. retrospective verbalization). As the student completed concurrent verbalizations for each of the items the researcher took notes of their reasoning. Later, after all the items had been responded to the researcher asked clarifying questions which resulted in the retrospective verbalizations. Each think-aloud session was video-recorded but the video recording was used to capture only what happened on or around the computer screen.

***Analysis: Coding of think-aloud interview data.*** All interviews were first transcribed, and transcripts included documentation of any gestures<sup>6</sup> or student working on paper. The think-aloud interviews typically lasted about 33 minutes each (min = 19, max = 51). During this time the student completed think-alouds for six items across the LT. On average, every item had data from four think-aloud interviews. Transcripts were coded using an open coding approach (Strauss & Corbin, 1990) by using the procedure shown in Figure 18. Each transcript was passed through the procedure to establish if the item elicited student thinking that was intended by the level/elaboration of the level. If the student(s) were using the kind of thinking as described in the

---

<sup>6</sup> Often, students would point at (using fingers or the mouse pointer) images on the computer screen or use gestures to explain their thinking about an item.

level, then the next step was to establish if the student was able to correctly answer the item. If the student was unable to answer the item correctly, it was important to check if issues such as poorly drawn images, level of readability, unfamiliar context, or the absence of a calculator were posing as hindrances. If the item elicited student thinking that was not intended by the level/elaboration but reflected a relevant/compelling mathematical practice for this group of students, it was important to document the strategies and verbalizations that students were using to answer the item. After reviewing the data from all the students who participated in a think-aloud for a given item, recommendations were made to: a) retain the item in its current state, b) make adjustments to the item, c) collect more data on an item, d) flag items at a level as having issues around opportunity to learn, or e) make adjustments to a level. Given the nature of the project, it was imperative that these recommendations were first put forth to the SUDDS learning sciences team before any final decisions were made. More importantly, the researcher needed to integrate these recommendations with the results from the quantitative analysis carried out to answer Research question 2.



*Figure 18. Procedure for coding think-aloud interview data for each item.*

## **Research question 2.**

*To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?*

Responding to this research question allowed the researcher to examine the structure of the items aligned to the LT levels by assessing the distribution of the item difficulty for each item. The aim was to determine whether the item difficulty increased in a monotonic fashion from level to level. In the event that it did not, a second aim was to identify items that were not conforming to the notion of monotonic increase in item difficulty and propose potential sources of variation.

**Data Source: Field-test data from MM 6-8 and Analyses from Think-Alouds.** The Similarity and Scaling LT has 23 field-tested items distributed across the 9 levels. Table 4 below shows the items associated with each of the levels and what type of item it is. These items appear on equated<sup>7</sup> test forms that typically have between 10 and 12 items and often take about 30-40 minutes for students to complete. Items marked with asterisks were used for think-alouds only, as they did not have a sufficient number of observed responses (n<100).

Table 4. Item associated with each level of the *Similarity and Scaling* LT.

Item ID	LT Level#	Item Type	No. of observations	Item ID	LT Level#	Item Type	No. of observations
749	1	MC	113	750	6	SM	134
1181	1	SM	194	771	6	NE	127
759	2	SM	164	773	6	SM	282
762	2	SM	190	800	6	SM	101
765	3	1L	268	751	7	NE	102
766	3	1L	125	770	7	NE	344
767	3	T/F	165	236	8	NE	137
763	4	MC	143	241*	8	MC	4
764	4	NE	147	433*	8	MC	5
768	4	NE	218	761	9	NE	223
760	5	NE	224	1182	9	NE	161
769	5	NE	193				

Key for Item Type abbreviations: MC – multiple choice, SM – select multiple responses, 1L – enter one letter (items that require matching), T/F – true or false, NE – numeric entry. Items 749, 763, 760, 769, and 770 are scored dichotomously, the remaining are scored polytomously.

**Analysis: IRT Rasch Models integrated with analyses from Research question 1.** The methodology used to analyze the data for a learning trajectory followed the model used in Confrey and Toutkoushian (under review) but was adapted to apply to a construct. Their model includes using a variable selection method to regression to first identify items that do not

<sup>7</sup> In the first year of field testing, forms were developed to align to both districts' scope and sequence documents and therefore, form equivalence across test forms associated with this LT was not possible.

conform to the structure of the LT. After identifying potentially non-conforming items, they describe a process for reviewing the item relative the three variation factors (construct irrelevant (IRRE-LV, inter-level variation INTER-LV, and intra-level variation, INTRA-LV).

Measuring a student's understanding of a construct can be estimated but not directly measured (Bond and Fox, 2015). From a psychometrics perspective, the simplest estimate of a student's understanding of a construct, often referred to in the literature as the latent trait, is their total score on an assessment. Similarly, the simplest estimate of the difficulty of an assessment item is the proportion of examinees who answered the item correctly. However, these estimates, which are artifacts of classical test theory (CTT) have limitations because they can only "interpret characteristics of people taking the test and of questions on the test in the context of one another" (Izsak & Templin, 2016, p. 6). That is, a test-taker's score on a test depends on the difficulty of the items on the test. Conversely, the difficulty of the items depends on the cohort of students taking them. To overcome this limitation, modern psychometrics offers item response theory (IRT) which involve a family of models that allow for interpretation of item characteristics, such as difficulty, independently of a specific group of test-takers (Izsak & Templin, 2016).

To begin with, the researcher carried out both single group IRT (Rasch) analyses and multi-group analyses (academic years as the groups) on the data to see whether the data from one academic year to the next needed to be analyzed separately. There are many ways of using fit statistics to determine which model fits the data better using (Templin, 2018). de Ayala (1998) suggests that lower AIC and BIC values imply better fit. Therefore, the models were compared and the fit statistics (AIC, AICC, SBIC, BIC and log likelihood) favored the single group Rasch model. The researcher then proceeded to use the simpler Rasch model to generate item

parameters for all 21 items. A Rasch model generates parameters for the item difficulty referred to as  $\beta$  parameters based on the number of possible points achievable. For example, Item ID 1181 includes two numeric entry parts, therefore, the model generates a parameter for getting one out of the two parts correct, and a parameter for getting full credit. For this study, the researcher chose to use the  $\beta$  parameter associated with getting all parts of an item correct so that these could be easily compared to the parameters generated for dichotomous items.

Once the item parameters were generated, the researcher used a simple linear regression to check whether there was a positive association between the LT level and the item difficulty parameters. The aim of this analysis was to find a linear regression model that generated a positive slope and a high  $R^2$  value, which would stand in as evidence of increasingly sophisticated student thinking as one goes up the LT. In addition to fitting a linear regression model, the other focus of the analysis was to systematically determine which items were depressing the regression (labelled non-conforming items) and needed further scrutiny from the learning sciences perspective.

An iterative variable selection method to regression process was applied to the data to identify non-conforming items for further scrutiny. First, a baseline regression model was run, then the item with the highest absolute residual value was removed, and the regression analysis was repeated. The new  $R^2$ , residual sum of squares (RSS), and slope were then compared to the baseline regression model. The ultimate goal was to maximize the  $R^2$  value while also retaining as more than 70% of the items, so the process of removing and comparing was repeated iteratively until the aforementioned criteria were met. Once the desired criteria were met, the Spearman rank correlation of LT level and the item difficulty parameter b for the remaining subset of items was calculated.

For the subset of items that were identified as non-conforming, the researcher used Confrey & Toutkoushian's (in review) process to scrutinize each item from a learning science perspective and determine whether the item was not conforming due to a) construct-irrelevant variation (Irrel-V), b) intra-level variation (Intra-LV), or c) inter-level variation (Inter-LV). At this stage of the analysis, the results of the regression were interpreted both quantitatively and qualitatively.

If an item had issues around readability, ambiguous diagrams, unfamiliar terms/contexts, being solvable by test-taking tricks, or other construct-irrelevant distractions such as too many steps, too many distractors, or too much time required to solve it, and the think-aloud data from *Research question 1* further substantiated these flaws, then the item was marked as Irrel-V. If items were different from other items at the same level because of factors such as numeric values, directness of the question, familiarity or ease with the representation, availability of a calculator, or availability of additional visual support, and data from think-alouds substantiated the presence of such factors then the item was marked as Intra-LV. If an item or all the items at a level are identified as showing Inter-LV, the item(s) is/are considered relative to other levels in the LT. If the item fits better at another level in substance and difficulty, the item(s) is/are moved.

In summary, a recommendation was proposed as to whether the item would be: 1) retired, 2) revised to address Irrel-V, 3) revised to address Intra-LV, 4) retained in the pool but post an alert for the users (teachers/students) to indicate that the item is known to be atypical (hard/easy), 5) retained but the description of the LT level would be edited to clarify range of or inclusion of items, 6) moved to another level (possibly as a set of items), or 7) used as evidence to rearrange the order of levels in LT.

### **Research question 3.**

*How do patterns of observed performance differ across students from a diverse sample of students on LT-based assessments measuring geometric similarity?*

Responding to this question allowed the researcher to examine whether the assessments aligned to the *Similarity and Scaling* LT in MM 6-8 were equally sensitive to the construct-relevant student thinking of a diverse sample of students.

**Data Source: Field-test data from MM 6-8 and Student Demographics.** These data sources have been described on earlier pages, specifically, in Table 3 and Table 1, respectively.

**Analysis: Comparing distributions of scores across ethnic and gender subgroups.** Typically, differential item functioning (DIF) analyses are used to answer variations of research question 3. The presence of DIF for a given item signifies that the item may be measuring something other than an individual's understanding of a target concept, i.e. what is supposed to be measured by an assessment item. DIF analyses can uncover items that may be potentially biased towards certain ethnic sub-groups (Martinková et al, 2017). However, due to the relatively small number of observations recorded for each of the items (per sub-group), it was not possible to analyze DIF for each of the 21 items. Instead the researcher generated distributions of total scores across all various ethnic groups and separately across both genders. Analysis of variance (ANOVA) was used to examine whether the means of each of the subgroups were significantly different from each other.

### **Research question 4.**

*How do the students' results on the MM 6-8 similarity assessment compare to those on an independent assessment measuring geometric similarity?*

Responding to this question allowed the researcher to examine whether scores produced by diagnostic assessments that measure LTs are correlated to scores produced by other assessments that measure student understanding in the equivalent domain.

**Data Sources: Similarity and Scaling LT assessment and Independent Assessment.** In addition to field test data for items associated with this LT, 136 eighth-grade students from District 1 took an assessment developed specifically for the study but was made up of a selection of the items in Table 4. Results from this assessment were used to select students for think-aloud interviews and was also used to carry out correlational analyses needed to answer the fourth research question. The assessment comprised of 8 equated test forms, each with nine questions (one from each level of the LT). The assessment was administered in Fall 2017 prior to students learning the 8<sup>th</sup>-grade geometry content on geometrical transformations.

*Independent Assessment.* District 1 subscribes to an assessment data bank developed by a well-known publishing house. With the permission of the school/principal the researcher derived access to a subset of items in the item bank (250 items) that were aligned to the same CCSS-M standards associated with the *Similarity and Scaling* LT. The researcher then mapped all the items to the progress levels of the LT (Table 5). Of the 250 items, none of the items aligned to the first four levels of the LT. In addition, there were 12 open-ended items and the remaining 238 items were all typical multiple-choice items. Open-ended items were not selected to maintain consistency in item types across the *MM 6-8* assessment and the independent assessment. To ensure fairness in terms of time required to complete the test, a 15-item test was created (Appendix I). The reason for this is that the assessment items from the independent assessment item data bank were typically shorter, and more procedural. In total, 241 students completed the written under test conditions. These tests were then hand-scored by the researcher.

Table 5. Mapping of Independent Assessment Item Bank to *MM 6-8* LT progress level.

Progress Level/Description	Number of Items
Level 5	3
Level 6	9
Level 7	216
Level 8	5
Poor Formatting	4
Constructed response item types	12
Misaligned to standard	1
Total	250

**Analysis: Correlational analysis of scores from both tests.** For a more meaningful comparison of both assessments because they were not generated using the same theoretical framework, it was imperative to first define ability parameters for each student based on their performance on each test using an IRT (Rasch) model. The Rasch model was used to generate ability parameters for each of the students based on their performance on the independent assessment items ( $\theta_{indep}$ ). Using the same model, parameters for each of the students based on their performance on the Similarity and Scaling LT assessment ( $\theta_{MM6-8}$ ) were also generated. The data from both assessments were merged so that the correlation of scores would be done for only the students who took both assessments (n= 131). The correlational analysis was then carried out on the ability parameters for each of the students based on their performance on the independent assessment items ( $\theta_{indep}$ ) versus the Similarity and Scaling LT assessment ( $\theta_{MM6-8}$ ).

## **Chapter 4: Results and Discussion**

This chapter is divided into four sections. Each section responds to one of the four questions and discusses implications of results in relation to the validity claim associated with that question. First, the results from the qualitative analysis of the think-alouds are reported. The discussion of these results is followed by a set of recommendations made to the SUDDS learning sciences team about a sub-group of items that the researcher found to be problematic after conducting analyses of the data. Next, a brief but pertinent discussion about what transpired from these recommendations is presented as they have implications on the results and discussion of succeeding IRT analysis. The IRT results establish a separate subgroup of items as non-conforming to the structure of the LT. These results are followed by an in-depth discussion from a learning sciences perspective of the items deemed non-conforming. In the third section, results from the comparison of score distributions across gender and ethnic subgroups are reported. Finally, results from the correlation between the *Similarity and Scaling* LT assessment scores and the independent assessment are reported.

### **Results from Think-Aloud Interviews**

#### **Research Question 1.**

*To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in MM 6-8?*

***Claim:*** *The items in the diagnostic assessments that measure students' understanding of geometric similarity are aligned to the corresponding LT levels as defined in the DLS MM 6-8.*

18 students from District 1 participated in think-aloud interviews conducted by the researcher. 23 assessment items spanning 9 levels of the Similarity and Scaling LT were utilized during these think-alouds. The think-aloud results have been organized by LT level. Each level description is

provided first, followed by screenshots of all the items aligned at that level, a brief description about how we expect middle grades students to respond to each item, and further, the kinds of systematic errors or misconceptions we expect to uncover. Then, a summary of response patterns from the think-alouds is provided to illustrate what was learned about the items. Each summary discusses whether the data from the think-alouds support the item(s) as sufficiently measuring the thinking as described in the level and other relevant issues e.g. noteworthy student thinking or strategies that came up during the think-aloud. Excerpts from the think-alouds are interspersed throughout the results section as evidence to support the researcher's stance on an item. Note that excerpts of interview transcripts are labeled with the student's think-aloud identification number e.g. TA 2, TA 7, TA 12 etc. Where there is a dialogue between the researcher and the students, the researcher is labeled with an R. Anything that the student or researcher refers to by pointing, writing on paper, or gesturing is captured in parentheses. The summary ends with a recommendation from the researcher to: a) retain the item in its current state, b) make adjustments to the item, c) collect more data on an item, d) flag items at a level as having issues around opportunity to learn, or e) make adjustments to a level.

***Level 1: Identifies two shapes as exactly the same or congruent by placing one on top of the other.*** There are two items at this level, Item 749 and 1181 (Figures 19 and 20). Item 749 is a multiple-choice<sup>8</sup> item which has been used as an example in the Materials and Tools section in the Methodology chapter. Item 1181 is similar to item 749 that assesses student misconceptions concerning the orientation or non-essential features such as color of a shape when informally checking for congruence. However, item 1181 is a different item type from item 749 because a student answering this item can only achieve full credit for selecting both the first

---

<sup>8</sup> In the first year of field testing (2015-16) Item 749 was a select multiple item and not a multiple choice

and third answer options. Selecting a correct response and an incorrect response will result in the student getting no credit for this item.

*Results: Item 749.*

Five students were discussing whether the two shapes below are congruent. Their comments are in the table below.

Student	Comment
Student 1	Shapes A and B are not congruent, because shape B is backwards.
Student 2	Shape A is congruent to shape B, because I can flip shape B and it fits exactly on top of shape A.
Student 3	Shapes A and B are congruent, because if I slide them together, they will fit into each other.
Student 4	Shapes A and B are not congruent, because they look different from each other.
Student 5	Shapes A and B are congruent, because they have the same area.

Which student is correct?

Student 1  
 [1]Student 2  
 Student 3  
 Student 4  
 Student 5  
 None of these students are correct.



*Figure 19. Item 749, Level 1.*

For multiple choice items such as this one, there is [1] next to the correct answer options.

Five students completed a think-aloud interview for this item. Three out of the five students answered Item 749 correctly during the concurrent verbalizations whereas the other two students found *Student 5*'s comment about the same area compelling and selected that as their final answer. During the concurrent verbalizations, none of the students considered the comments that stated the shapes were not congruent (*Student 1* and *Student 4*'s comments). For example, one student said:

**TA 8a:** Shape A and B, I believe they are congruent cause they look the same shape just flipped around. So, I'm going to cross out Student 1, um and Student 4. Yeah student 1 and student 4.

Four students contemplated *Student 5* as a plausible answer, but only two of the four followed through and selected that as their final response because they believed that

measurement or the act of finding the area were mathematically better than simply “flipping”.

Their reasoning is shown below:

**TA 12:** 2 and 5 are rational to me. I mean this one (Student 2) is true cause they do. I pick Student 5 because it has stuff to do with more mathematics rather than, “Oh I can flip it”. [I]t’s a more mathematical thing, cause if it’s talking about they have the same area which is true, because they’re still the same shape but it’s mirrored around, so it’s still the same measurements just in a different placement. But Student 2 is also correct because they are congruent because Shape B is just a flipped version of Shape A. But I also realized well congruent means that they have the same measurements, so I was like that goes with area, so I chose Student 5.

**TA 8b:** I honestly would go with 2 more, but I think that 5 would be the right answer. Because it is talking about area. And it is geometry, so I would just want to go with 5.

For the other two students who were contemplating between *Student 2* and *Student 5*, one reasoned that because the flip would allow the shapes to fit exactly on top of each other, she would pick *Student 2* as her final response:

**TA 5:** I was doing student 5 also, because it looks like they do have the same area, cause they’re the same exact shape but they’re just flipped. So, I was thinking that, but then, when it flips on top of each other it will actually like fit on top of shape A, so that’s why I was like it’s going to be student 2 not 5.

On the other hand, the second student felt forced to respond with *Student 2* as her final answer because there were no measurements provided for her to find the area:

**TA15:** But measurements will tell you say, these two (points out two vertical sides of the trapezoids) or these two (points out the slant sides of the trapezia) are the same. They would just tell you, it’s just backwards. But you can’t really tell because you don’t really have measurements. Then I would assume that student 2 was correct because there’s not a lot of information and you have to go with what you think is right.... Well 5 is gone because we don’t have any measurements, so we don’t know if they have the same area.

Only one student explicitly referred to *Student 3*'s comment during her concurrent verbalization, allowing the researcher to probe for further explanation during the retrospective verbalization:

**TA5:** If you slide it, these sides right here (vertical side on shape B) will go over here (points to left side of shape A) but you want them to match up so that's why you flip (gestures flip). But when you slide it won't fit together because this (left side of shape A) is slanted and this is (left side of shape B) is straight so, it will go a little bit more out. And then this will leave like a space right there (gestures that right most side of shape A would be uncovered).

*Results: Item 1181*

Consider the two triangles below.

Triangle 1      Triangle 2

Which statements about triangle 1 and triangle 2 are true? Check all that apply.

- [1] make triangle 2 fit perfectly on top of triangle 1.
- Triangle 1 is not congruent to triangle 2 because triangle 2 is upside down.
- Triangle 1 is congruent to triangle 2 if they have the same side lengths and angle measurements.
- Triangle 1 is not congruent to triangle 2 because they are different colors.
- Triangle 1 is congruent to triangle 2 if they have the same area.

*Figure 20.* Item 1181, Level 1.

For multiple select items such as this one, there is [1] next to ALL correct answer options.

Five students completed a think-aloud for Item 1181. Two students answered it correctly during the concurrent verbalizations by picking only the first and the third answer options. One of them had begun by selecting the last answer option during the concurrent verbalization but talked herself out of selecting it during the retrospective interview. She began by saying that:

**TA 7:** Because area, focuses on length, how big the triangle is. Since they are the same size, then if they're the same area then they're the same triangle.

During the retrospective verbalization, the researcher asked her if it was possible to have triangles that have the same area but not look the same, and the student said it was possible but when it came to her drawing them out on paper, she was unable to provide specific dimensions that would give the same area for different looking triangles:

**R:** Okay so, do you think you could find two triangles that have the same area but don't look the same?

**TA7:** Yeah (deselects last answer option).

**R:** Can you show me an example of one?

**TA 7:** Could be like...this triangle (draws a right triangle) a right triangle and then an equivalent triangle. Or like an isosceles triangle.

**R:** Okay, so you're saying these two are congruent, but they have different areas?

**TA 7:** No.

**R:** They have the same area?

**TA 7:** (nods) But they're not congruent.

The other student also attempted to provide a numeric example during the retrospective interview, she attempted to provide different dimensions that would yield the same area but at the time, was unable to provide numbers that would support her claim:

**TA 14:** Okay, so even though it's the same area it doesn't mean it has the same dimensions. Like, um, you want it to be the same and equal, it has to have at least a ratio for the dimensions. And if it's an area you could say that this would be...

Two of the students selected only the third and last answer options while one of the students (TA 18) selected the first, third, and last answers options. TA 18's reasoning highlighted that the answer options that included an "if" logic statement were perhaps not being read or understood as intended. For example, she reasoned:

**TA 18:** And if they have the same side lengths and the same angle measurements I don't think they can be a different shape. And if you can make them perfectly fit on top of each other then they must be the same area.

*Discussion of Level 1 items.* Level 1 describes how students informally check for congruence between two shapes by placing them on top of another and checking to see that one shape completely maps on to the other. From the think-alouds it is evident that both items 749 and 1181 are sufficiently measuring student thinking as highlighted by this level description. However, placing the two shapes that are thought to be congruent on top of one another to check for same shape and same size and knowing that if they match they must have the same area is quite different to recognizing that even though two triangles have the same area they are not necessarily congruent. This is a distinction that is nonetheless important to make even when informally learning about congruent shapes. Data from the think-alouds suggest that the notion of triangles being congruent because they have the same area is very compelling to some of these students. Furthermore, the notion of “necessary” versus “necessary and sufficient” conditions for testing for congruence are not formally discussed in the 7<sup>th</sup> grade<sup>9</sup> CCSS-M standard associated with this LT. Therefore, it is a possibility that students may not have had explicit instruction related to congruent areas being a necessary but not sufficient condition for testing the congruence of shapes. The researcher recommends that both items should be flagged so that teachers can be made aware of this particular issue. Teachers can explore the notion of area as a necessary but not sufficient condition informally in the middle grades by using Dynamic Geometry Environments such as GeoGebra® or Sketchpad ® and asking questions such as: is it possible to create two triangles that “look the same” but do not have the same area, or is it possible to create two triangles that have the same area but do not “look the same?”

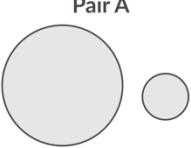
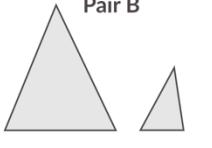
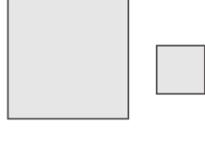
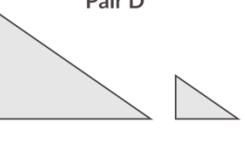
---

<sup>9</sup> Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale.

**Level 2: Identifies similar objects as having the “same shape” and justifies by informally scaling, e.g., centering or matching vertices.** There are two items at this level: item 759 (Figure 21) and 762 (Figure 22). Both items are written to allow students to informally assess similarity by working with properties of shapes such as vertices. Item 759 looks at similarity across different types of shapes and not just triangles. In order to score full credit on item 759, students must select all three marked options (circle, square, and right-angle triangle). On the other hand, item 762 checks for students’ misconceptions such as an overgeneralization of similarity of triangles e.g. all right-angle triangles are similar. To gain full credit on item 762, students must select only one answer option.

*Results for Item 759.*

Below are four different pairs of shapes.

<b>Pair A</b> 	<b>Pair B</b> 
<b>Pair C</b> 	<b>Pair D</b> 

In which of these pairs could the larger shape be shrunk to become congruent to the smaller shape? Check all that apply.

[1] Pair A  
 Pair B  
 [1] Pair C  
 [1] Pair D

*Figure 21. Item 759, Level 2.*

Five students completed a think-aloud for Item 759. All five students who did a think-aloud for this item received full credit by selecting all the correct answer choices. They typically used language that signified informal scaling when explaining why the square and circle will shrink to become congruent to the smaller shape. Informal centering was not possible to observe

as the shapes are not movable on the screen. As a result, their retrospective verbalizations tended to be along the lines of the examples given below:

**TA 12:** They're both squares, if you were to shrink the square eventually you would probably get the same shape.

**TA 5:** Pair A can because the circle, you shrink it, it will just turn into a smaller circle.

When deciding on the triangles in Pair B, students often began their reasoning by classifying both triangles. TA 5 and TA 15's comments below about the "slant" illustrate how both these students looked for deviations from prototype:

**TA 5:** I thought that pair wasn't a congruent shape because to make a triangle smaller it would stay the same but instead you have the big triangle and then you have this slanted kinda triangle and when you make this (larger) triangle smaller it will just stay the same and not become more slanted.

**TA 15:** This (larger) looks like equilateral and that one looks like it has a slant.

One of the five students appeared to be using a strategy that by-passed any reasoning about the properties of the shapes and instead appeared to be "spotting the odd one out" from the four pairs of shapes:

**TA 8a:** And after that I would glance over all of them. And I would notice that Pair B does not seem similar. While the other ones did, so I selected those.

*Results from Item 762.*

Carl wants to know which of the smaller triangular shapes could be enlarged on a photocopier to be congruent to the corresponding larger shape.

Pair A	Pair B
Pair C	Pair D
Pair E	Pair F

In which pairs, if any, are the triangles the same but different sizes? Check all that apply.

- Pair A
- Pair B
- [1] Pair C
- Pair D
- Pair E
- Pair F

*Figure 22. Item 762, Level 2.*

Five students completed a think-aloud for Item 762. All students picked the correct answer choice for this item i.e. Pair C as being two shapes that would be congruent if the smaller triangular shape was enlarged using a photocopier. Four out of the five were able to articulate the reason for their choice and their reasons were alike in that they used a form of informal scaling by using gestures to indicate they were pulling the hypotenuse of the smaller triangle towards the larger triangle:

**TA 16:** C was just how this (mouses over “hypotenuse” of the smaller darker triangle). I was looking at the shape of the right triangle, a right triangle always has this (mouses over right angle vertex) and just the way how if this kinda nudged upwards (mouses over top vertex of smaller darker triangle). It will fit almost perfectly, because of how it already fits perfectly over here (mouses over shared vertex). And how every other side looks like it would be equal. (Mouses over gap between two hypotenuses) So basically basing it off the shape and enlarging it in my mind to see how it would look.

In addition to picking the sole correct answer choice, all the students picked at least one of the other, incorrect choices. The most common choice among the five students was Pair B. The retrospective verbalizations indicated that they were attempting to enlarge the shapes by

using gestures that they would use on touch devices or in software applications (e.g. enlarging an image in MS Word). Below are excerpts from TA 9 and TA 11's think-aloud interviews, and their responses suggest that both these students are not identifying similar triangles as having matching vertices:

**TA11:** Like if we enlarge in that one (C) would just go the left (gestures dragging hypotenuse of smaller triangle towards the hypotenuse of the bigger triangle). So, that one (B) is going upwards (gestures dragging top vertex of small triangle towards large triangle's top vertex). So that's why I feel like in my head they fit together.

**R:** What about Pair B?

**TA9:** Um, cause I saw it as if you could like enlarge it (makes a smartphone-like gesture of flicking finger apart) it would be the same.

**R:** So which way would you enlarge it?

**TA9:** This way (points along an imaginary line between the top vertices of both triangles).

*Discussion of Level 2 items.* Level 2 describes students' strategies such as informally scaling or checking to see if the angles at the vertices match. Overall, data from the think-alouds suggests that both items address the LT level sufficiently but there are two separate issues at play. The first issue is that it is possible to answer item 759 using a "spot the difference" kind of logic as TA8a demonstrated. This implies item 759 is also measuring construct irrelevant thinking and should therefore be revised. A recommendation is made to adjust the answer options to remove the redundancy between the circle and square and further, include a pair of right triangles that are orientated differently from each other. The latter ensures the item captures any student misconceptions while also ensuring that a student is not using the aforementioned "spot the difference" logic to gain credit for this item.

The second issue has to do with how item 762 is scored by the system. It is very likely that students are thinking correctly about the pairs of triangles in item 762 but the nature of the select multiple item type is disallowing many students from getting any credit for that thinking. While scoring is an issue, based on the think-alouds it is evident that this item sufficiently addresses the level. Furthermore, it elicits misconceptions such as “all right-angle triangles are similar” or over-generalization such as “all isosceles triangles are similar.” There is some redundancy with two of the pairs: D and E. The researcher recommends replacing either D or E with a new pair of triangles that are similar to allow for more opportunities to get partial credit for this item but still maintain the integrity of the item’s alignment to its level.

***Level 3: Identifies similar triangles as having matching angles but possibly differing side lengths.*** There are three items at this level: 765, 766, and 777. Items 765 and 766 (Figures 23 and 24, respectively) both provide a pool of different triangles from which students are expected to find pairs of similar triangles. On the other hand, item 777 (Figure 25) provides four pairs of triangles, each with a similarity statement and students must decide whether the statement is true or false based on the information provided in the diagram. In each of the three items, students must respond correctly to all parts of the item to gain full credit. A student who is simply matching a single pair of angles across the two triangles has a misconception and would, therefore, pick any two right-angled triangles as being similar without checking to see if the other two pairs or even one more pair of angles was the same. Students who have persisting ideas about same orientation will pick triangles that are oriented in the same way even though they do not have matching angles.

*Results: Item 765.*

Ten triangles are shown below.

For each of the following triangles, enter the letter of a similar triangle, other than itself. If no other triangle is similar to the given triangle, enter N.

Triangle A [1] [F]	Enter a letter
Triangle B [1] [E]	Enter a letter
Triangle C [1] [G]	Enter a letter
Triangle D [1] [H]	Enter a letter

*Figure 23. Item 765, Level 3.*

For 1 letter items such as this one, the correct answer is written next to the [1].

Five students completed a think-aloud for Item 765. Out of the five students who did a think-aloud for this item, three selected the correct string of answers: F, E, G, and H. One student selected F, J, G, and H thus getting three out of the four parts correct. Another student selected N, J, G, and H thus getting only two out of the four parts correct. All five students picked the correct answers for the last two parts of the item because these were typically straightforward choices and none of the students contemplated any of the other distractors in the pool of triangles (A – J). The three students who correctly answered all parts of the item explicitly matched up the three angles in the given triangles with one that was in the pool of ten triangles. They did so during the concurrent verbalizations by calling out the three angles when they saw them and used the mouse cursor to point them out. During the retrospective verbalizations, the researcher asked the students why it was important for the numbers to be equal in order for the triangles to be similar. One response indicated that the student was informally matching the vertices on the screen to show that the triangles “looked the same”:

**TA 3:** I think this one tells you how long it is from this side (points to 20 in triangle A) and I think the 70 tells you how it is on the top. And they're like matching (points to 70 and 20 in triangle F). Cause like the smallest one is at 20, like where the point ends is at 20 (gestures an angle with index finger and thumb) and where it starts at is 70. And it's doing the same with this one (triangle F) and so that's why I was like Oh, it looks the same.

Two students approached it using some variation of an informal transformation of one triangle to the other:

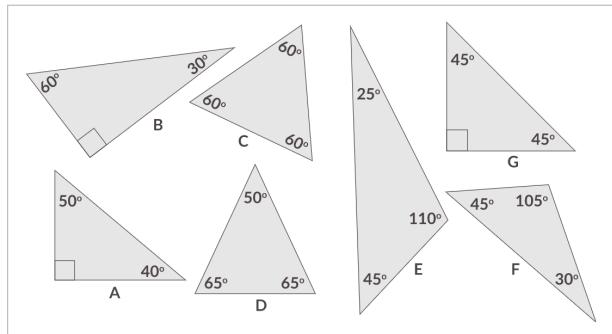
**TA 13:** I can't show you. Um,...kind of like you can take all triangles and you can like, I don't, I don't know,...yes, you can like, even though they're like not all the same sizes or the way like they sit, facing that way and that's facing (shows the varying orientations of triangles)...even though they're not the same size they're the same because they have the same angles. So, like you could get this triangle (points to a triangle) from that triangle (points to a triangle).

**TA 17:** Hmm. In my mind I was like you can say that it's smaller like that's (A) smaller than that (F) but then in my mind I was like flipping the triangle (F) and so it was matching that one (A) but that one's (draws triangle) just a tiny bit smaller. But they're both right angles and they both have the same angles.

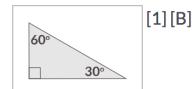
Only one student (TA 6) was referring to the size or approximate areas when she made her choices for all but the last part in the item. Her retrospective verbalization indicated that she thought that none of the triangles in the pool matched Triangle A because "if you look at the structure and size of the triangles there was no triangle that was similar to triangle A." When asked to elaborate on how she looked for same structure she replied, "Like the same size of it and the similarities to it." She provided analogous reasoning for why she picked Triangle J as being similar to Triangle B: "They are both similar and I feel like this one (J) is a little bit smaller but they are both the same."

*Results: Item 766.*

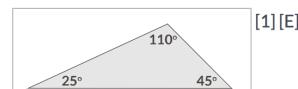
Seven triangles are shown below.



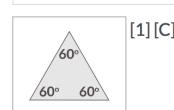
For each of the following triangles, enter the letter of a similar triangle.



Enter A, B, C, D, E, F, or G



Enter A, B, C, D, E, F, or G



Enter A, B, C, D, E, F, or G

*Figure 24. Item 766, Level 3.*

Six students completed a think-aloud for Item 766. Four out of the six students correctly answered this item. Like the students who did the think-aloud for item 765, they called out the three angles during the concurrent interviews as they were matching them. The other two students had varied responses. One of them believed that if the authors of the item wanted the numbers (angles) to be the same, they would have used the word “same” and not “similar”:

- R:** You didn’t pick C because it had exactly the same numbers? And so, your idea of similar is that it can’t be exactly the same?
- TA2:** Well, if it wanted to be it would have said find the one that looks the “same” as this one.

In contrast, the other student believed the item authors were referring to the classification labels of triangles as a definition of similarity:

- R:** So, when you’re saying type of triangle, you’re saying they could have been the same because this (triangle in the first part) and these three (triangles G, B, and A) are all right triangles? And this (triangle in second part) and these two (triangles E and F) are...what types of triangles are they?
- TA7:** Obtuse or acute?
- R:** And this one you said was an equilateral triangle? (answer option 3)
- TA7:** And there’s no other equal triangle.

- R:** Correct me if I am wrong, you're wondering if it has to do with the type of triangle or the measures of the triangle. What do you think it is?
- TA7:** I think it is the type of triangle because that would, like, be too obvious. And so, like this (90, 30, 60 triangle) could be G or A (deletes B and tried to enter both G and A).

*Results: Item 767.*

Below are four pairs of triangles and statements about them.

Pair	Diagram	Statement
1		$\Delta ABC \sim \Delta FED$
2		$\Delta ABC \sim \Delta ADE$
3		$\Delta ABC \sim \Delta DEF$
4		$\Delta ABC \sim \Delta DEF$

Indicate whether the statement in the table about each pair of triangles is true (T) or false (F).

Pair 1 [1][t]

Enter T or F

Pair 2 [1][t]

Enter T or F

Pair 3 [1][f]

Enter T or F

Pair 4 [1][t]

Enter T or F

*Figure 25. Item 767, Level 3.*

Five students completed a think-aloud for Item 767. Four out of the five students answered all parts of item 767 correctly. Given that the students did not have to pick out from a pool of triangles (unlike Item 766), most students generated responses rapidly by checking to see if all angles in both the triangles matched.

Only one student did not read the statement in the second pair of triangles and thought that it was a comparison between a triangle and a trapezoid: “It’s false because the shapes, it’s not cut right. Like here(BC) it’s not the same type of shape. It’s a trapezoid and a triangle.” When asked what the symbol,  $\sim$  meant, many of the students did not really recognize it. This is likely because the formal treatment of similarity happens in 8<sup>th</sup> grade. They used words such as “equal”, “approximately equal”, and “about”. To further investigate how the students were

making their decisions for triangles that did not have the same angle measures, the researcher asked them how they knew they were different:

**TA 6:** Because this one (ABC) is taller like thinner, and this one (ADE) looks a little flatter like bigger like that (gestures with both hands)

**TA3:** I can just look at it and say that it's false. I say that because how they're shaped. One is more, like the angles are more pointed it's like going up, and the other one is just like a regular triangle but just going sideways. So, I would say that's false. And for this one (equilateral) it's true. All they did was try to turn it upside down to see if it would like trick you. So, I think it's true.

TA 3 seems like could also be using the “same shape but different size” argument to support her reasons for not selecting the third pair of triangles. However, she does refer to the pointedness (size) of the angles.

*Discussion of Level 3 items.* Even though similarity is not formally covered until 8<sup>th</sup> grade, students have an intuition for checking to see how shapes “look the same” by checking the angles at the vertices. Some student reasoning with items 765 and 766 and the patterns of responses suggest that picking a triangle from a pool of triangles requires students to focus on angles at the vertices and not on orientation, size, or type (classification) of triangles. Therefore, reasoning required to answer items 765 and 766 will need to be more nuanced than that for item 767 where students are given a pair and have to decide whether the pair is similar or not. However, both approaches i.e. picking out similar triangles from a pool and checking to see if a given pair is similar are both foundational to a formal treatment of similarity and provide for rich classroom discussion about the meaning of geometrical similarity in the case of triangles. All three items are aligned to the level and the only recommendation for change (even though it did not hinder students from responding to the item) would be to have a brief note on item 767 explaining what the symbol  $\sim$  means.

*Level 4: Grows or shrinks a similar triangle by doubling, tripling, or halving side lengths and defines similar triangles as having congruent angles and proportional sides*

There are three items at this level. Item 763 (Figure 26) is written with the assumption that students recognize that in order avoid distorting a shape when re-sizing it, the dimensions of the shape must be scaled in proportion. Students who select the second and the fourth answer options in this item may be using additive reasoning instead of multiplicative reasoning. On the other hand, students who select the last answer option may be using correct multiplicative reasoning but not within the context of the question. Item 764 (Figure 27) uses the context of constructing a ramp for a porch that is similar in slope to an existing ramp but is required to reach double the height of the existing ramp. Additional questions are asked to check that students recognize that the base ratio of both the ramps are the same even though they are different in height and base. Item 768 (Figure 28) uses the context of triangular pieces that can be used as base building blocks to build larger triangular shapes of any size. Students are then expected the scale up from the base triangular piece to work out the dimensions of larger triangular shapes built from it.

*Results: Item 763.*

Diedre wants to put the logo below on the back of her cell phone. She measures the logo and finds that it is 8 inches long by 6 inches wide. Her cell phone is 6 inches long by 3 inches wide.

Diedre knows that she needs to shrink the logo for it to fit on her phone. Which size logo will fit on her phone and not be distorted?

- 2 in. long by 3 in. wide
- 5 in. long by 3 in. wide
- [1] 4 in. long by 3 in. wide
- 6 in. long by 4 in. wide
- 6 in. long by 4.5 in. wide

*Figure 26. Item 763, Level 4.*

Three students participated in the think aloud for item 763 and only one student used proportional reasoning to come up with the correct answer. During the concurrent verbalizations she eliminated the last two options as she stated they were too wide for the phone, she eliminated the 5 in. by 3in. option stating that it would be “fuzzy”, and when asked why 2 in. by 3in. was not a correct option gave the following reasoning during the retrospective verbalizations:

**TA 16:** It didn't make sense to me because usually, when you try to make something smaller and didn't want it distorted, it means you usually grab it by one side (mouses over top left corner of the square in which the image is embedded), and push it in together (motions with mouse some sort maintaining aspect ratio motion) as to where it is depleting by the same thing each time and each side is doing the same thing. So, when it said two by three, there was no way to get a 2 here (points to vertical dimension) if you get a three on this side. So, I stuck with 4 times 3. Because I know three times two is six and four times two is eight. So, they're both depleting, they're both going down by um, well, they're both diving by two.

The other two students had varying responses but had not understood what the question was asking of them. One of them attempted to use the Pythagorean Theorem stating that he was trying to find solve for a missing side. The other student did not know the word distorted meant

and went on to pick the 5 in. by 3 in. option. Similar to the student who reasoned correctly, he eliminated the last two options by stating that the logos were too wide for the phone but picked the 5 in. by 3 in. as it would be more prominent than the 2 in. by 3 in or the 4 in. by 3 in: “You want a bigger logo so everybody can see it. So, I picked 5 and 3 was the width. That’s why I chose 5.”

*Results: Item 764.*

Stuart's grandmother uses a wheelchair, and needs a handicap ramp to enter Stuart's house. Stuart's porch is 40 inches tall. Stuart wants to build his grandmother a handicap ramp that is the same steepness as the one at the hospital. The hospital's ramp has a vertical brace every foot. The hospital ramp's dimensions are shown below.

vertical braces

20 inches

10 feet

vertical braces

How far out into his yard would the ramp go (in feet) if it starts at the top of the porch and matches the steepness of the hospital ramp?

[1]  $20.0 \leq x \leq 20.0$

Enter a number

When a person in a wheelchair goes up the hospital ramp, each time she passes a brace, how many inches higher off the ground is she?

[1]  $2.0 \leq x \leq 2.0$

Enter a number

When a person in a wheelchair goes up Stuart's ramp, each time she passes a brace, how many inches higher off the ground is she?

[1]  $2.0 \leq x \leq 2.0$

Enter a number

*Figure 27. Item 764, Level 4.*

Six students completed a think-aloud for Item 764. Two out of the six students answered all parts of this question correctly by using proportional reasoning during the think alouds:

**TA 14:** Okay, so forty inches tall. So, I have multiply that by two and everything by two for dimensions. So, if we're going end up multiplying twenty by two we have to multiply ten by two so, it would be twenty (enters 20 into answers field).

One student, TA 18 applied proportional reasoning but assumed that the number of braces did not change i.e. Stuart's ramp also had 10 braces and thus did not get the correct answer to the last part:

**TA 18:** Okay that one was two and Stuart's ramp has to go up 40 inches not 20 so if they go up the ramp each time would be 4 feet? (re-reads question) This one she's two inches off the ground so that one she's 4.

From the three students who did not use proportional reasoning, one did not understand the context of the question and abandoned the question, one tried but had much trouble understanding the question, and one reasoned correctly about the structure of the braces using repeated addition. Her reasoning went along the lines of:

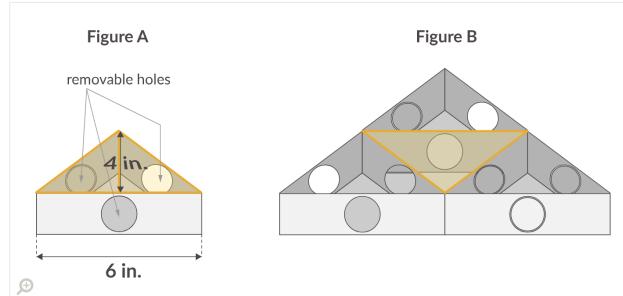
**TA 5:** So, I was thinking, like I can get 4 and then 4, 4, 4, 4, (shows iterating hops using fingers) and that should equal the 20.

However, she was using every other brace which indicated that she had not entirely understood the image in relation to what the second part was asking.

*Results: Item 768.*

Tina invented a new way to build hamster cages. She designed a triangular plastic shape that could snap together with identical shapes to build larger hamster cages.

The plastic shape has a triangular top with a base of 6 inches and an altitude of 4 inches (see figure A below). Each plastic shape has three rectangular walls with removable circles to let the hamsters move between the plastic shapes. Using four plastic shapes, she built the large triangular hamster cage shown in figure B below.



How wide is the base of the large triangular top of figure B?

[1]  $12.0 \leq x \leq 12.0$

Enter a number

What is the altitude of the large triangular top of figure B?

[1]  $8.0 \leq x \leq 8.0$

Enter a number

Tina wants to make a larger triangular hamster cage with a 42-inch wide base.

How many triangular plastic shapes should she place along the base?

[1]  $7.0 \leq x \leq 7.0$

Enter a number

What would be the altitude of the top of this larger triangular cage?

[1]  $28.0 \leq x \leq 28.0$

Enter a number

How many triangular plastic shapes would this larger triangular cage have in total?

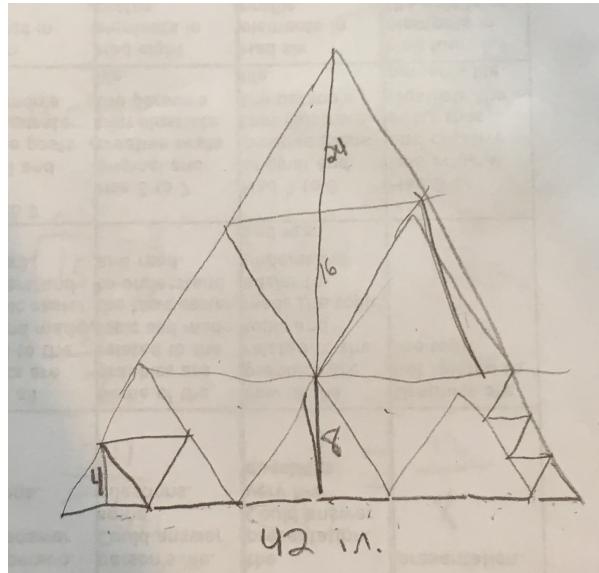
[1]  $49.0 \leq x \leq 49.0$

Enter a number

*Figure 28. Item 768, Level 4.*

Five students completed a think-aloud for Item 768. None of the students were able to gain full credit in this question during the think-alouds however, three out of the five were able to correctly answer the first three parts of the question. They used either repeated addition or

scaling to generate their responses. These three students used some form of a drawing (base of seven triangles) to respond to the last two questions but did not create complete diagrams and thus ran into problems with miscounting (Figure 29).



*Figure 29.* An example of a student's incomplete drawing used to answer item 768.

None of the students used scale or proportional reasoning to generate an answer for the altitude of the triangle with a base of seven triangular pieces. One out of the five students (TA 6) used the image in the item and repeated addition to work out the answers for the first two parts but was unable to go further even though she attempted to draw it out. Figure 30 shows TA 6 attempt at drawing out the cage with seven pieces at the base but she thought she had to draw it out exactly as it was in the item i.e. with holes and other details.

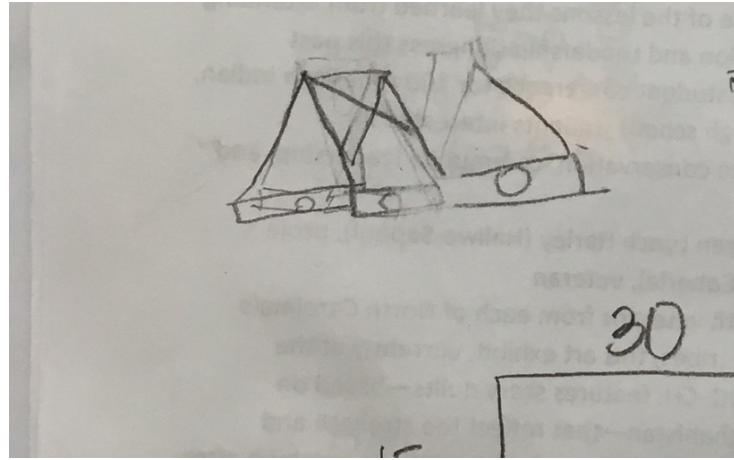


Figure 30. TA 6's drawing of the hamster cage.

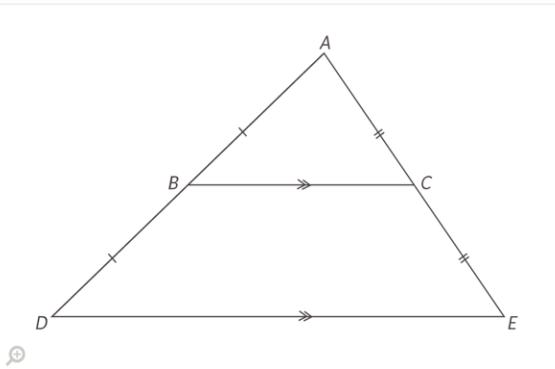
*Discussion of Level 4 Items.* At a minimum, all of the level four items have the capacity to elicit whether students are maintaining proportional relationships when growing or shrinking triangles. They also help discriminate the students who may be using additive reasoning from those that are beginning to use proportional reasoning. However, the think-alouds highlighted some issues with all three items. Firstly, item 763 has the word “distortion” which will need to be explained with a footnote or replaced with a kid-friendly phrase that better conveys the idea of maintaining the proportions of an image e.g. *Which size logo will fit on her phone and not be stretched or squashed?* For students who had trouble reading, deciphering the text in the stem of the question to the images in item 764 (hospital ramp) and item 768 (hamster cage) was a challenge. Therefore, it is recommended that the amount of reading is condensed, and the diagrams are better labeled without making the question overly easy e.g. label the whole image as “hospital ramp” and label the ramps. Additionally, the main text of the question needs to explicitly state that both ramps have the same structure i.e. braces at every foot. The image of the hamster cages could be altered so that they are shown to be upright, so that student can tell that the base is what the cage is “resting” on and the altitude is how tall it is. Furthermore, the last part of item 768 (hamster cage) asks students to find the number of triangles it would take to

make up a cage that has a base of 42 inches. This prompt requires functional reasoning (e.g. input → output) which none of the students demonstrated (those who got to part 4 used an image and miscounted). This implies that this part of item 768 is not well aligned to the LT level/elaboration and should be removed.

***Level 5: Defines similar triangles as having equal angles and sides in proportion and, given one pair of corresponding sides, solves for unknown corresponding sides.*** There are two items at this level item 760 (Figure 31) and item 769 (Figure 32). Both items have two similar triangles given and students are expected to use the sides in proportion to solve for an unknown side. Item 760 does not expect student to use AA criteria to establish similarity prior to finding the missing side length i.e. the students are already told that both triangles are similar in the stem of the question. Item 769 leverages the context of shadows that line up at a given time as means to calculate the height of tall objects such as trees.

*Results: Item 760.*

B and C are the midpoints of DA and EA, respectively. BC is parallel to DE. Triangle ABC is similar to triangle ADE.



If the length of DE is 18, then what is the length of BC?

[1] [9.0 ≤ x ≤ 9.0]

Figure 31. Item 760, Level 5.

Three students completed a think-aloud for Item 760. All three students who completed think-alouds completed this item correctly. Only one of the students did not use proportional reasoning during her concurrent verbalizations:

**TA7:** So, BC looks about half of DE, so divide 18 by 2. Which will make it 9. The arrows, if you look at the arrows and then, they're the same right here (points to arrows showing parallel lines) but then this one (DE) is still longer. So, it represented half to me.

The other two responded with reasoning that reflected that they understood that the scale factor between the two triangles was 2:

**TA 14:** B and C are midpoints of D and A. Ok, yup, that's in between, and EA. B is in between D and A and C is in between E and A. Okay, so if the length of DE is 18, then what is the length of BC? Okay, that's a midpoint. So, if it's the middle, scaling down by 2. I wanna say 9 (enters 9 into the answer field). Let me see. DE, it's in the midpoint so yah, nine.

### Results: Item 769

Clara's foot-long ruler makes a 16-inch shadow on a flat surface. At the same time, a tree lined up with Clara's ruler makes a shadow that is 120 feet long, as shown in the diagram below. Clara can use this information to determine the height of the tree.

NOT DRAWN TO SCALE

How tall is the tree in feet? [1]  $[90.0 \leq x \leq 90.0]$   
Enter a number

Figure 32. Item 769, Level 5.

Five students completed a think-aloud for Item 769. None of the five students who participated in the think-alouds for this question gained any credit for this item. Only one student indicated that she was thinking about a proportional relationship between the length of ruler's shadow and the length of the tree's shadow but did not take into account the different units of measurement:

**TA8b:** I figured if you, if one foot would be sixteen inches based on the shadow then 120 feet, I thought that if you divide 120 feet by 16 then you would get how tall it is because it's when it says one foot it is like counting how tall it is and not what it laying down flat. So, I was like, if I divide 120 by 16 then I'll get however tall it is.

Among the remainder of the students, one attempted to use the Pythagorean theorem:

**R:** You said something about Pythagoras?

**TA 8a:** The Pythagorean theorem?

**R:** Yeah, the Pythagorean theorem. Tell me why you were thinking about Pythagorean theorem?

**TA 8a:** Cause, it's finding um, missing lengths of a triangle. But it was a while back when we went over that. So that's why I was having struggles on this problem.

While the other student was comfortable disregarding the "not to scale" note on the image:

**TA 11:** I was saying over here it says 1 foot (point to image). So, I was just going upwards with it. (iterating 1 foot up on the image)

**R:** Upwards to where?

**TA 11:** Like to where the outline right here is (gestures extending dotted line across image on the screen). For the height cause that's what we're trying to find. So, I just added within that and it looks like it would be 3 more than that one.

**R:** Can you see that information that's in gray over there? What does that mean?

**TA 11:** Not drawn to scale, like, it's not perfect.

**R:** But you're still comfortable with doing just the 3 more?

**TA 11:** Uhuh.

*Discussion of Level 5 items.* Item 760 which looks at two embedded triangles that are in proportion (without a context) appears to be relatively easy for students but it is probable that students are responding with the correct answer simply because the length of the larger triangle “looks” like it is half the smaller triangle. To ensure that the item is eliciting proportional reasoning about the triangles, an extra part could be added to the item which asks the students to find ratio of the pair of corresponding side lengths (other than the base).

Conversely, the context of shadows in item 769 may be unfamiliar for many students with an added confusion around the different units of measurement. The AA criterion for similarity is not covered in 7<sup>th</sup> grade, and so it is not clear if many would reason using the two objects (tree and ruler) that are at 90 degree-angles with the flat ground. The item should be revised so that the information about the two triangles being similar is explicit in the stem of the question. This edit may enable more students who are not familiar with the context of shadows to access the item. Additionally, item 769 can be better aligned to its LT level by reducing any bias associated with the conversion of units.

***Level 6: Extends similarity relationship from triangles to other shapes, and identifies similar shapes with proportional corresponding sides and congruent corresponding angles.***

There are four items at this level. Item 750 (Figure 33) presents a pool of different rectangles and students are expected to use proportional reasoning to pick a rectangle that is similar to a given rectangle. To gain full credit students must select only one correct rectangle. Selecting any of the other rectangles would indicate that the students are using additive reasoning. Item 771 (Figure 34) presents two similar trapezoids with one pair of corresponding sides given and the students are expected to find three missing side lengths using proportional reasoning. Item 773 (Figure 35) uses a context of enlarging photos without distortion. Students who respond by selecting the

second and third answer options are using additive reasoning. Students who only respond by selecting the last two options are able to work with integer scale factors to identify similar rectangles and students who select all three correct answer options show the flexibility to work with fraction scale factors. Item 800 (Figure 36) is about a parent who wants to enlarge a rectangular soccer field in a manner than maintains the ratio of its width to length. Students who are only using integer scale factors will only select the answer option that is twice the width and the length of the original soccer field. While students who are considering fractional scale factors will select both the correct answer options i.e. twice the original dimensions and  $1\frac{2}{3}$  times the original dimensions.

*Results: Item 750.*

Below are six rectangles that are different sizes.

 A	 B	 C
 E	 D	 F

Which rectangles are similar to rectangle D? Check all that apply.

- A
- B
- [1]C
- E
- F

*Figure 33. Item 750, Level 6.*

All four of the students who attempted this item selected C based on the fact that its side lengths were half the lengths of the corresponding sides of rectangle D. However, all of them selected one or more other rectangles. TA 15 was unique in her approach; she applied correct proportional reasoning with her choices but made a calculation error when working out the ratios of side length of rectangle D and E:

**TA 15:** Cause 4 times 1.5 is 6 and 6 times 1.5 is 8. And then for C, I don't think it's A  
(deselects A) I don't think dilation works with subtraction just multiplication and division so. And then D to C is 4 divided by 2 is 2 and 6 divided by 2 is 3.

Two of the students selected C but also selected A and E. TA 3 used visual inspection while TA 7 used additive reasoning:

**TA 3:** This one's a little more ... (C) I wouldn't say long but no, C can go... I know B can't because this one more tiny-er (gestures with fingers to indicates narrowness of rectangle) than longer. A looks, more longer on top than this one (D). It can look similar, but it looks more skinny-er. I would say, E, A, C.

**TA 7:** Well because the rest didn't have the same but. E, I guess it could have been E too or A because A you could have subtracted 1 and E you could have added 2.  
(changes her mind and selects A and E).

When TA 7 asked why she did not select these options to begin with i.e. during the concurrent verbalization she responded in a manner that indicated that she was recalling a rule learned in class:

**TA 7:** Because I thought all of them had to be divide by 2 or multiply 2, or something with 2. Yeah, because it could be, as long as you can do something to both of the numbers to get the other numbers then yeah.

*Results: Item 771.*

In the figure below, trapezoid ABCD is similar to trapezoid AGJE.

The length of each side of trapezoid ABCD. If necessary, round your answers to one decimal place.

The length of side AD is \_\_\_\_ [1]  $1.6 \leq x \leq 1.6$   
Enter a number

The length of side BC is \_\_\_\_ [1]  $3.6 \leq x \leq 3.6$   
Enter a number

The length of side AB is \_\_\_\_ [1]  $2.8 \leq x \leq 2.8$   
Enter a number

*Figure 34. Item 771, Level 6.*

Five students participated in think-alouds for this item and two of them demonstrated proportional reasoning. One of these students was unable to move forward with the question because 2 did not “go” into 5:

**TA 8a:** I am trying to see if I can find a way like how I can use the lengths on the bigger shape  $AGJE$  to figure out the lengths of  $ABCD$ . Two doesn't go into five and that throws me off a little bit. If they're congruent they should have similar side lengths too. That would compare and then you could figure out the other side lengths too.

The other student was initially distracted by the quadrilaterals  $DEFC$  and  $CFJH$ . After finding contradictory lengths for the segment  $CF$  using her visual inspection she started over at which point she reasoned that:

**TA 16:** It has to scale it down to this cause it's equal, similar actually. It's five, let's see how many times it goes into two. So, two goes into five (works out on paper) add an extra place, ... it would have to be two point five.

She used the long division algorithm and the scale factor of 2.5 to find all the missing lengths for the smaller trapezoid. The other three students attempted to find the missing lengths

using a variation of visual inspection e.g. dividing  $AE$  into two, as it “*looked that way*” in order to find the missing side  $AD$ , “ $BC$  should be less than 9”, or halving the length of  $AG$  to get the missing length  $AB$ .

*Results: Item 773.*

Trevor has a photograph of his grandmother that is  $4'' \times 6''$ . He wants to enlarge it to fit in a larger frame with no white space around the photograph and no distortion.

Which frames could he use? Check all that apply.

$5'' \times 7''$   
 [1]  $6\frac{2}{3}'' \times 10''$   
  $8'' \times 10''$   
 [1]  $8'' \times 12''$   
 [1]  $10'' \times 15''$

*Figure 35. Item 773, Level 6.*

Four students completed a think-aloud for item 773. Only one student selected all three correct answers, but her reasoning was mathematically inaccurate:

**TA5:** I started thinking to choose these three because I saw that, these two number (4 and 6) are even I was like maybe all the numbers are even maybe that they will still fit. And then with this one ( $6\frac{2}{3}$ in. by 10in.) I was just thinking because there's already a 6 (points to stem), it probably won't hurt to make the 6 a little bit bigger, and then make it bigger with the 10. So that's why I chose that one.

The same student, when asked what distortion meant to her responded with the following:

**TA 5:** I would think that distortion is something like no spaces in between so it fits perfectly. That's what I would think.

One of the students would have selected all three correct answers but when dividing 10 by 6, she rounded her final answer to 1.6 instead of leaving it as a fraction or mixed numeral. When she multiplied the other dimension of the photo (4) by 1.6 it did not match the answer

option  $(6\frac{2}{3})$ . The other two students only selected the 8 in. by 12 in. as the final answer and neither of them knew what the word distortion meant: “probably no space I guess?” or “It means that some of it won’t miss. Like some of it won’t disappear.”

*Results: Item 800.*

Arthur painted boundaries for a rectangular soccer field in his backyard for his kids. The dimensions are shown below.

A diagram of a rectangle with a white border on a dark green background. The bottom edge of the rectangle has a label "length = 30 ft." and the right edge has a label "width = 15 ft.". There is a small magnifying glass icon in the bottom-left corner of the rectangle.

Arthur chose these measurements when his children were small, but now they are bigger. They asked him to make a bigger soccer field that has the same ratio of length to width as the one to the left. Which of the following measurements could he use to make this bigger field? Check all that apply.

- Length = 30 ft. and width = 30 ft.
- Length = 60 ft. and width = 15 ft.
- [1] Length = 60 ft. and width = 30 ft.
- [1] Length = 50 ft. and width = 25 ft.

*Figure 36. Item 800, Level 6.*

Four students answered this question during the think-alouds. Two out of the four students used proportional reasoning but only one of them selected both correct answer choices. One of the students used repeated addition ( $30 + 30$  and  $15 + 15$ ) to find one of the correct answers but did not select any other response because:

**TA 6:** ...in my perspective there wasn't another answer solution. So, I chose that one.

And lastly, the most interesting response was that of a student who selected both correct responses but used an elimination strategy:

**R:** You picked 50 and 25 first. Can you tell me more about why you chose them first?

**TA 10:** It's cause he's trying to make it better.... bigger! And if it was 30 again it would still be small. So, that wouldn't work. And if it was 50 it would have been a little bit bigger for them. And if it was 25 – it would have been about the right shape.

**R:** You picked that one (60 and 15) and then you didn't pick it.

**TA 10:** It's cause of the width, if this (points to length) was longer than this (points to width) has to be longer too. It wouldn't have been the right shape – it would have been (gestures a wide shape).

**R:** So how did you know that 50 and 25 would make it about the same shape?

**TA 10:** I mean it would still make it a shape if it was used by 60 and 15 but it wouldn't make the, like, a little bit right you know? You know like the three little bears and goldilocks?

**R:** And tell me why you picked 60 and 30?

**TA 10:** Oh, it can't be the same number again, cause what's the point in even fixing it!

TA 10 was able to pick both correct answers without using any proportional reasoning whatsoever which implies that the other answer options are not doing enough to distinguish students who are using test strategies from those that are using proportional reasoning.

*Discussion of Items at Level 6.* All four items at this level elicit a variety of reasoning associated with the level description; however, based on the think-alouds there some proposed edits to three of the items. Item 750 elicited some rule-based reasoning such as “what you do to one you must do to the other” together with some informal reasoning such as “it looks like it” rather than proportional reasoning. It would be interesting to see student responses if a rectangle that was proportional to the given rectangle ( $D$ ) but oriented differently was added to the pool of rectangles in a revised version of this item. The image in item 771, as it stands, may be confusing for a number of students who feel compelled to find relationships where there are none. A proposed change would be to not embed the trapezoids and have them sit side-by-side in the image. Item 773 fits and is able to discriminate between students who are working with integer value scale factors from those that are comfortable with fractional scale factors. Additionally, it can also expose students’ additive misconceptions. A minor language edit proposed to item 773 is to replace the word “distortion” with a word/phrase that students in 7<sup>th</sup> grade may be more familiar with such as “enlarged so that it looks the same and is not stretched or squashed.” Item 800 also discriminates between students who are working with integer value scale factors from

those that are comfortable with fractional scale factors but one of the students' reasoning when eliminating the incorrect responses indicates that the item may need a new answer option that is larger in terms of both side lengths than the original shape but also not in proportion e.g. a 45 by 30 rectangle.

**Level 7: Uses similarity, proportionality, and scale factors to solve problems involving scale drawings.** There are two items at Level 7: item 751 (Figure 37) and item 770 (Figure 38). Both these items are typical of what is found in classroom materials that address the 7.G.A.1 CCSS-M standard (i.e. Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale.). Item 751 uses the context of a floor plan to check if students can scale down by a factor of  $\frac{1}{12}$ . Item 770 uses the content of scales on maps to check if students can find the actual distance between objects represented on a map.

*Results: Item 751*

Chantelle's mom is remodeling the kitchen in her home. The actual dimensions of the kitchen are shown below. The designer she hired brought a scale drawing of the new kitchen for Chantelle's approval before construction.

The drawing is a  $\frac{1}{12}$ -scale model of the real kitchen. Complete the following sentences.

The actual length of the kitchen is 15 ft. The length of the kitchen in the scale drawing is \_\_\_\_ ft.  
 Enter a number

The actual width of the kitchen is 12 ft. The width of the kitchen in the scale drawing is \_\_\_\_ ft.  
 Enter a number

The actual area of the kitchen is \_\_\_\_ times as big as the area of the kitchen in the scale drawing.  
 Enter a number

Figure 37. Item 751, Level 7.

Five students completed a think-aloud for Item 751. During the think-alouds none of the students were able to answer all three parts of this item correctly and one student was unable to answer any of the parts. Three out of remaining four students were able to find the correct answer for the actual width of the kitchen i.e. 1 foot, but only one was also able to find the correct answer for the actual length of the kitchen. An interesting example was that of a student who went around in circles during the concurrent verbalizations because he felt compelled to set up a proportional statement. Initially, he reasoned correctly and entered 1.5 into the first part but later he changed his mind. He stated that he had to set up a proportional statement. As he set up the proportional statement on paper, it was evident that he was not comparing corresponding sides and thus ended up with 180 feet for length of the actual kitchen. He claimed that this answer was in inches and followed up by dividing it by 12 just to end up where he started i.e. at 15 feet. His final response pattern was 15, 12, and 0. During the retrospective verbalizations, the researcher asked him not to worry about setting up a proportional statement but to explain what he thought the question was asking him to do. Below is an excerpt from the conversation that followed:

- R: [L]et's not worry about setting up the proportion. Tell me what that sentence means to you – the drawing is one-twelfth scale model of the real kitchen. In your own words, what does that mean to you?
- TA 8a:** That the scale model is only 1/12 the size of the actual model
- R: And what does that mean? Is it bigger? Is it smaller? Is it the same?
- TA 8a:** Smaller (without hesitating)
- R: Why do you say smaller?
- TA 8a:** I guess it's cause 1/12 is a smaller fraction plus it's not one whole. So, I am going to have to assume it's smaller.
- R: But you are not sure about the answers you put in because?
- TA 8a:** They're the same, they're the same as the actual length. Actually, for the actual kitchen let's find the area. Area equals base times height so, 12 times 15, (does working out on paper). Well, there is no difference since I got the same measurements. So maybe, actually that doesn't make sense. I think I'll put zero in, since it's the same.

Two out of four students used visually inspections of the length of the image on the computer screen to find their answer to the scaled length of the kitchen (2 and 3) but used correct reasoning to work out what  $\frac{1}{12}$  of 12 feet was (answer to the second part). Neither of them was able to answer the last part of the item.

*Results: Item 770.*

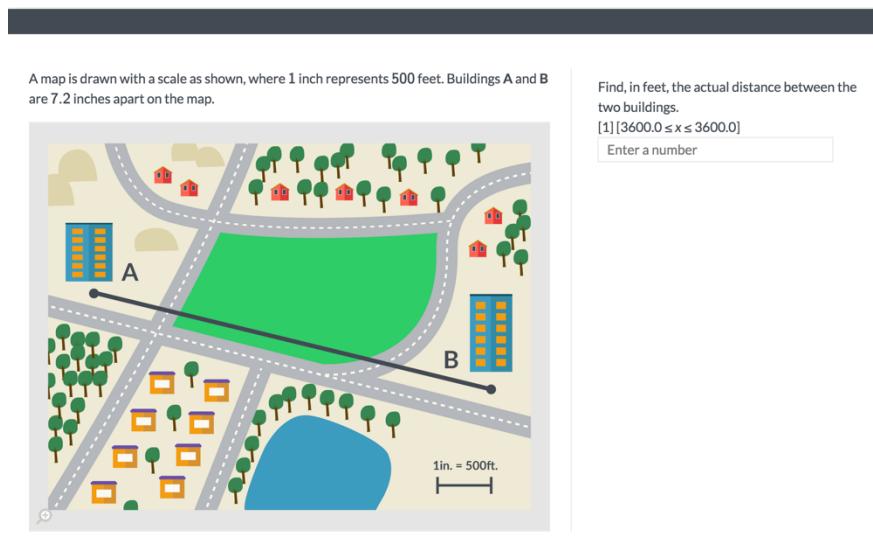


Figure 38. Item 770, Level 7.

Six students completed a think-aloud for Item 770. Four of the students multiplied the scale factor by the map distance to generate their answer using either the multiplication algorithm or the splitting the 7 and the 0.2 of an inch, turning the .2 of an inch into a rational number and then multiplying it by 500. Only one student stated explicitly during her concurrent verbalization that she was multiply the two numbers because the scale on the map represented a proportional relationship between the scaled distances and the actual distances. Others tended to state that they “had” to multiply the two numbers. One student responded in the following manner when asked why it was:

- R:** Tell me a little bit more about why you did that? (point to the working out on paper)
- TA 9:** I saw that 1 inch equals 500 feet. And so I took 7.2 inches and times it by 500.
- R:** And why did you know that times-ing was the right thing to do?
- TA 9:** Cause when you add it, I mean it doesn't really make sense. I learned that it's not accurate when you add them. Inches to feet you can't add those two, you multiply them.

Two of the students did not answer this question correctly. One of the students did not know where to begin with attempting the question. Whereas the other lined up part of her pencil with the scale on the map, and then approximated the number of iterations (6) of the portion of the pencil that would get her from *B* to *A* on the screen. However, she was unable to transition to what that meant in terms on actual distance.

*Discussion of items at Level 7.* Item 751 appears to measure this level more completely than item 770. Item 751 has a minor issue which is that the scale factor as well as the conversion factor between feet and inches is the same. TA 8a's think-aloud illustrates the confusion that can ensue from this overlap. TA 8a's think-aloud interestingly points out that contexts are important in helping students think about the reasonableness of their answers. It would be interesting to see what his response would have been in a context free question had he setup an incorrect proportion i.e. would he have questioned his working and said that 180 must be a measure in inches?

Item 770 on the other hand is a typical textbook style question providing students a context that is very commonplace i.e. maps. It aligns to the LT level in its current form but appears to very loosely measure student understanding described at this level. For example, TA 9's response provides evidence that students could be multiplying two numbers without knowing why multiplication worked. One way of revising item 770 would be to add a second part to the question which asks students to justify their answer. To enable automated scoring, this part could

be list four or five justifications (e.g. to convert 7.2 inches to feet you have to multiply by 500, or every inch on the map represents 500 feet in the real world) from which the student would be expected to select the one that they believed explained their response in the first part.

***Level 8: Describes the effect of scaling a linear unit on area and volume.*** There was only field-tested item at this level, item 236. Given that all the other levels in the LT had at least two items, the researcher selected two items that were still being piloted, in addition to item 236 for the think-alouds. Item 236 (Figure 39) uses the context of 3 different tile sizes that have side lengths in proportion. Students are expected to extend their understanding of scaling a linear unit to work out what the pro-rata cost of a proportionally smaller and proportionally larger tile size. Item 241 (Figure 40) provides the dimension of a net of a rectangular prism and states that the dimensions are going to be doubled, it asks students to work out what the relationship between the volumes will be (2, 4, 6, or 8 times the original prism.) Item 433 (Figure 41) is similar to 241 but asks for the relationship between the areas when the net of a square-based pyramid is double in dimension.

***Results: Item 236.***

A hotel owner is replacing the tile in his ballroom. There are three sizes of tile to choose from, listed below.

Tile dimensions	Price
6 in. × 6 in.	?
12 in. × 12 in.	\$1 per tile
18 in. × 18 in.	?

How much would one 18 in. × 18 in. tile need to cost, in dollars, for the cost per square inch to be the same as the 12 in. × 12 in. tiles?  
[1]  $2.25 \leq x \leq 2.25$   
Enter a decimal number

How much would one 6 in. × 6 in. tile need to cost, in dollars, for the cost per square inch to be the same as the 12 in. × 12 in. tiles?  
[1]  $0.25 \leq x \leq 0.25$   
Enter a decimal number

*Figure 39. Item 236, Level 8.*

Five students completed a think-aloud for Item 236. None of the students were able to answer this item correctly during the think-alouds. A typical response during the concurrent verbalization was to establish a linear scale factor between one of the side lengths with another:

**TA 8b:** Ok, since, the first thing that would come to my mind is six is half of twelve, so if it's one dollar and that's half then I would divide it by two to get 50, so already that would be 50 so, I'd just add to that one and I would get 1.50. So, I think that would be the answer for the first one.

The researcher asked two out of the five students about how many 6 in. by 6 in. tiles would be needed to cover the space that a 12 in. by 12 in. tile during the retrospective verbalizations. In both cases, students stuck with their answer which described the effect of scaling areas the same as scaling lengths. The most frequent response for the second part of this item was \$0.50 with varying responses for the first part: 1.25, 2.50 and 1.50.

*Results: Item 241.*

Jen needs a box to hold a cake she is baking. She cuts a piece of cardboard into the shape below and folds the sides to make an open box.

The box is not large enough to hold the cake, so Jen decides to use a cardboard piece in the same shape, but she doubles each of the current dimensions. By doubling the dimensions of the cardboard box, how has Jen changed the volume of the box?

- The volume of the new box is 2 times the volume of the old box.
- The volume of the new box is 4 times the volume of the old box.
- The volume of the new box is 6 times the volume of the old box.
- [1] The volume of the new box is 8 times the volume of the old box.

*Figure 40. Item 241, Level 8.*

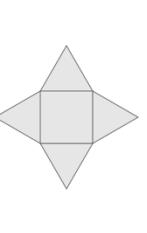
Six students completed a think-aloud for item 241. Two out of the six students actually worked the volume of the two boxes and then found out how much bigger the volume of the new

box was compared the original box by diving. None of the other four students used this method, nor did they use any fact-based relationship to answer this question. There were a variety of responses and not the typical misconceptions that students have, which is that they expect changes in area or volume to be directly proportional to changes in side lengths. It is possible that this handful of students may not have had the opportunity to learn this content, given their varied responses.

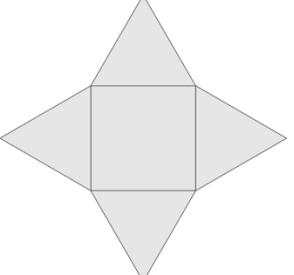
*Results: Item 433.*

Chardae bought a rectangular pyramid to use as a piñata at a party. She wants to make a bigger piñata that looks exactly the same as the original, so she doubled the length of all the edges of the pyramid. Then she cut a new net out of cardboard with these new measurements. Both nets are shown below.

Original piñata net



Enlarged piñata net



How much more cardboard will the enlarged piñata use than the original piñata?

- 2 times as much cardboard
- [1]4 times as much cardboard
- 8 times as much cardboard
- There is no way to know without the measurements of the piñatas.

*Figure 41. Item 433, Level 8.*

Five students completed a think-aloud for item 433. Four out of the five students claimed that it was not possible to find out how much cardboard will be needed without knowing the exact measurement of the piñata:

**TA 7:** I think it would have to be “There is no way to know without the measurements of the piñatas” because you don’t have any numbers to work with on each of them and you can’t really look at it and guess.

Only one student indicated that she thought the area of the cardboard required would be twice as much as the original piñata:

**TA 15:** Well she's making,... she's doubling the lengths, so probably two times as much (selects 1st answer option) seeing as she doubled the length.

*Discussion of items at Level 8.* All three items appear to be measuring the thinking required at this level of the LT including the types of misconceptions that highlighted: a) one must have the measurements to establish the relationship between the two areas/volumes, or b) that changes in area or volume are directly proportional to changes in side lengths. Therefore, no specific recommendations or changes are offered for the Level 8 items.

***Level 9: Uses informal arguments about angles to establish the AA criteria for similarity and solves similar triangle problems in context.*** There are two field tested items at this level: item 761(Figure 42) and item 1182 (Figure 43). Item 761 uses the context of town planning, parallel streets (lines) and street intersections (intersecting lines) to get students to think about establishing similarity on a more formal basis e.g. by referencing the angles made by parallel lines or intersecting lines. Students are given the dimensions of a larger triangular shape and expected to establish similarity and subsequently find the lengths of the missing sides of the smaller triangular shape. Item 1181 follows a similar structure but is within the context of two objects that are moving away at the same velocity from a common point.

*Results: Item 761.*

An architect is designing a park for a small triangular piece of land formed by the intersection of Broadway Avenue and Main Street. She needs the dimensions of the piece of land. She knows the dimensions of the larger triangle, as shown, and she knows the length of the park along Broadway Avenue is 13 meters.

What is the length of the park along Main Street? [1]  $[12.0 \leq x \leq 12.0]$   
Enter a number

What is the length of the park along Elm Street? [1]  $[5.0 \leq x \leq 5.0]$   
Enter a number

*Figure 42. Item 761, Level 9.*

Five students completed a think-aloud for Item 761. Three out of the five students answered this question correctly during the think-aloud. However, it was clear from concurrent verbalizations that students were not necessarily establishing AA criteria for the two triangles. Instead, they were using informal mapping and visual inspections to establish that the two triangles are similar to each other which was confirmed during the retrospective verbalizations. Here are two examples:

**R:** You worked out 12 and 5. How did you know to relate 65 and 13 to begin with?  
**TA 14:** Scaling. So this one (smaller triangle) is just a scaled down and flipped over so if I were to flip it over, flip it, and scale it up (motions using fingers the transformations) 13 would actually be 65.

**R:** But you were pointing at this (25 m) one specifically. What made you think that that was somehow connected to that (Elm)?  
**TA 3:** Because of the way it is lined up. Because that's (25 m) the back of this and that's (Elm) the back of this one.  
**R:** So you're saying that (25 m) is that back of this triangle (larger) and this (Elm) is the back of this triangle (smaller)  
**TA 3:** It can't be half of it – cause I was going to say that but then this is (Elm) way more smaller than that (25 m) ... I think is one (Elm) might be like 5 meters cause

if you put it like here like line up and there will be five of those. (enters 5 into answer key)

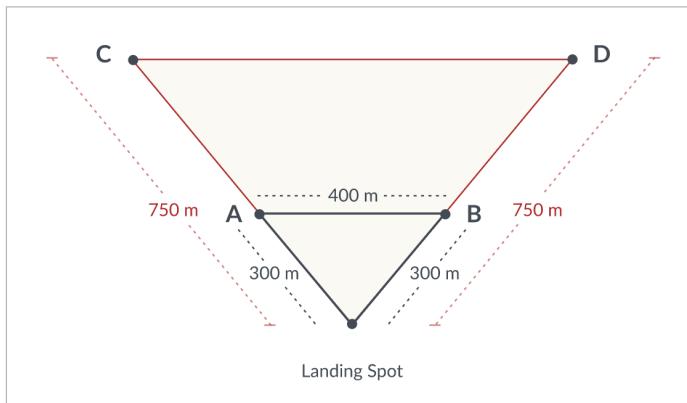
Only one student out of the five came closer to using the thinking described by the level during the retrospective verbalizations; she used her informal matching of vertices to establish that the smaller triangle was similar but just flipped version of the larger triangle:

**R:** Um, so were there anything in the image that confirmed to you that it was a flipped version of that bigger triangle? What about the image sort of drew you to that conclusion?

**TA 14:** I'd say how the, ... this is a ninety-degree angle (points to Main and Elm vertex), this is a ninety degree angle (point to 25 and Main vertex). So basically this angle right here (points to Main and Elm vertex) would be this angle (points to 25 and Main vertex). This angle (vertically opposite angle formed by intersection on larger triangle) is as widened up as this angle here (vertically opposite angle formed by intersection on larger triangle) right here where these dotted lines are. And this angle (angle made by Broadway and Elm) is as widened out as this one over here (angle made by Broadway and 25). So, basically, the lines really helped with that.

*Results: Item 1182.*

Two rovers are exploring a planet in a nearby star system. They start at a landing spot and drive away in straight lines at the same speed. When the rovers are both 300 meters away from the landing spot at points A and B, they are 400 meters apart from each other. They continue driving in straight lines, as shown in the figure below.



Calc

When the two rovers are at points C and D (750 meters away from the landing spot), how far apart (in meters) are they from each other?

[1]  $1000.0 \leq x \leq 1000.0$

Enter a number

When the two rovers are 12,000 meters apart from each other, how far away are they (in meters) from the landing spot?

[1]  $9000.0 \leq x \leq 9000.0$

Enter a number

Figure 43. Item 1182, Level 9.

Five students completed a think-aloud for Item 1182. Four out of the five students were able to answer the first part of this item correctly. Like Item 761, two of the students were using informal scaling to find their answer but two of the students indicated that it was a dilation and they need to find a scale factor. None of the students specifically established the AA criteria before proceeding with their calculations. TA 4's reasoning was very interesting and showed that he was able to respond to the item by using informal scaling:

**TA 4:** Oh then I read it right. It's about 800.

**R:** What made you think it was 800?

**TA 4:** This side this is 400 meters so you push up it get you to another 4,...that's 800 meters. And make it longer so That's about a 1000.

**R:** Okay it's about a 1000?

**TA 4:** Yeah.

**R:** Why do you think 1000?

**TA 4:** Because, if you push it up twice that's 800. If you do that triangle up here that's about 800. Then you got part of it still there. So, you like maybe shift the triangle over. Get you to 200 or 100, that would get you to a 1000.

On the other hand, the two students who were using the term dilation during their concurrent verbalizations, when asked what made them think that the smaller triangle formed by the landing spot,  $A$ , and  $B$  was a dilation of the larger triangle (landing spot,  $C$ , and  $D$ ) also reasoned informally:

**R:** Why don't we go back to the last one. I'm going to ask you one very simple question. Can you tell me why you thought this was a dilation?

**TA 13:** Because they're both triangles. And I knew that you could multiply something by 300 to get 750.

**R:** Anything else?

**TA 13:** The triangle (points to triangle formed by  $A$ ,  $B$  and landing spot) was inside of the other triangle (points to triangle formed by  $C$ ,  $D$  and landing spot). So, you could just increase this (motions to space between  $AB$  and  $CD$  to show an informal transformations approach.)

*Discussion of items at Level 9.* Level 9 is written to measure students' informal arguments about angles to establish the AA criteria for similarity which they then apply to similar triangle problems in context and solve for missing side lengths. Both items at this level,

749 and 1182 appear to be measuring only part of the thinking described at the level. That is, students are able to solve for missing sides using informal scaling and not the properties of the figures such as parallel line angles or angle sum of triangles to establish similarity between the two triangles. These two items do sufficiently measure the kind of thinking that is described lower down in the LT (Level 5: *Defines similar triangles as having equal angles and sides in proportion and, given one pair of corresponding sides, solves for unknown corresponding sides*).

### **Summary of Item Analyses from Think-Alouds.**

Table 6 shows an overview of how the 23 items were classified based on the think-aloud data. The researcher has added an additional column next to the classification column because an item was classified as *ADJUST* based on a variety of reasons (see Figure 18 for coding procedure) and not necessarily that the item did not measure the level. Where the item deviated significantly from the level, and the researcher had specifically discussed that the item or parts of it did not measure the level completely the researcher marked the recommendation with an asterisk (\*). After examining the transcripts and compiling the results for each item, four specific recommendations were made to the learning sciences team at SUDDS:

- a) Remove Level 1 items from any further analyses but retain them in the item pool and flag them so that teachers can be made aware of the potential Opportunity to Learn issues. Teachers can incorporate a more fuller treatment of congruence e.g. facilitating discussions about necessary conditions versus sufficient conditions before proceeding on to geometric similarity.
- b) Move Level 9 (*Uses informal arguments about angles to establish the AA criteria for similarity and solves similar triangle problems in context*) items to Level 5 (*Defines similar triangles as having equal angles and sides in proportion and, given one pair of*

*corresponding sides, solves for unknown corresponding sides*) as they are not completely aligned to the level and a student can respond correctly to the items without necessarily using the kind of thinking proposed by Level 9.

- c) Move Level 9 to be between Level 5 and 6 (*Extends similarity relationship from triangles to other shapes and identifies similar shapes with proportional corresponding sides and congruent corresponding angles*) and write new items that are better aligned to the level description. The level description will also need to be clarified to ensure that students are using reasoning related to angle sum of triangles, parallel line angles, etc. to establish similarity. An example of the new level description can be: *Uses properties such angles sum of triangles and parallel line angles to establish the AA criteria for similarity and solves similar triangle problems in context*. Additionally, Level 9 was placed at the top of the LT because students do not cover the necessary prerequisite knowledge e.g. angle sum of a triangle until 8<sup>th</sup> grade as per the CCSS-M (Confrey, 2015, personal communication). However, many researchers (Seago, Jacobs & Driscoll, 2010; Usiskin, 2014) have promoted a transformational approach to similarity which they claim would enable students to gain a more robust understanding of the concept. Unfortunately, due to the structure of the CCSS-M, most students do not have the opportunity to leverage transformations of shapes in order to better understand similarity when it is first taught i.e. in 7<sup>th</sup> grade.

Table 6. Classification of items based on data from think-alouds.

Item	Level	Classification	Recommendation
<b>1181</b>	1	FLAG	Collect more data
<b>749</b>	1	FLAG	Collect more data
<b>759</b>	2	IRRE-V	Make changes to pairs of shapes*
<b>762</b>	2	ADJUST	Remove a pair of triangles
<b>765</b>	3	FITS	NA
<b>766</b>	3	FITS	NA
<b>767</b>	3	FITS	NA
<b>763</b>	4	ADJUST	Revise the word “distortion”
<b>764</b>	4	ADJUST	Improve diagram and question stem clarity
<b>768</b>	4	ADJUST	Improve diagram and question stem clarity, Remove last part *
<b>760</b>	5	ADJUST	Add part to check for proportional reasoning
<b>769</b>	5	ADJUST	Improve question stem clarity
<b>750</b>	6	ADJUST	Change orientations of rectangles*
<b>771</b>	6	ADJUST	Improve diagram clarity
<b>773</b>	6	FITS	Revise the word “distortion”
<b>800</b>	6	ADJUST	Include another credible distractor*
<b>751</b>	7	ADJUST	Change fractional scale factor
<b>770</b>	7	FLAG	Add part to check for proportional reasoning*
<b>236</b>	8	FITS	NA
<b>241</b>	8	FITS	NA
<b>433</b>	8	FITS	NA
<b>1182</b>	9	PARTIAL FIT, MISPLACED LEVEL	Move to Level 5, Rewrite and Relocate Level 9*
<b>761</b>	9	PARTIAL FIT, MISPLACED LEVEL	Move to Level 5, Rewrite and Relocate Level 9*

### Summary of Student Understanding of Geometric Similarity from Think-Alouds.

Overall data from the think-aloud interviews confirmed for the most part that the items corresponded to the ways of thinking described by the levels. Level 1 (*Identifies two shapes as exactly the same or congruent by placing one on top of the other*) revealed an important characteristic of reasoning about similarity that had not been previously discussed in the literature i.e. students incorrectly believe that if two shapes have the same area, then they must be congruent. In terms of the learning trajectory, which is meant to inform instructional decisions, this can be handled in two possible ways. An implicit way of addressing this is to

follow the recommendation by the researcher about leaving the Level 1 assessment items in the item pool and flagging them for teachers to become aware of students' misapprehension that while congruence ensures equal area, equal area does not ensure congruence. An explicit way of addressing this issue is through elaborating Level 1 further to include one more misconception flag within the level description. Figure 44 shows the current misconception associate with Level 1. A second misconception flag could read: "Believes that two shapes with the same area must be congruent."

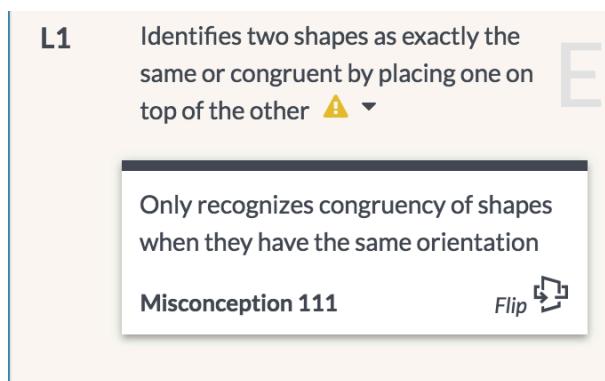


Figure 44. Current misconception associated with Level 1 of the *Similarity and Scaling LT*.

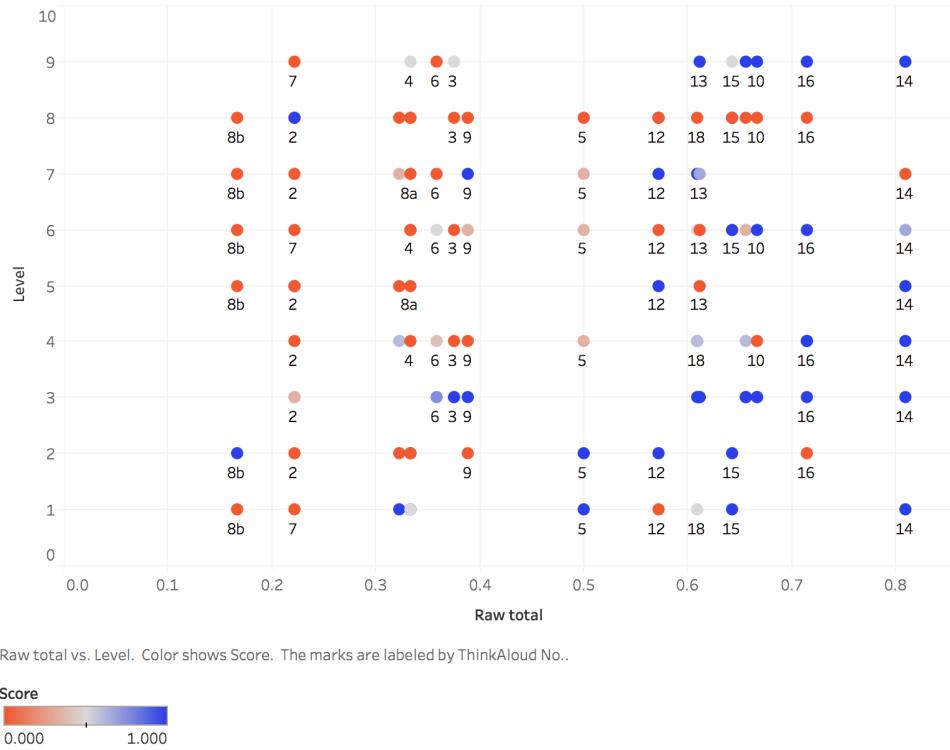
One of the items at Level 3 (*Identifies similar triangles as having matching angles but possibly differing side lengths*) also revealed an important characteristic of reasoning about similarity that had not been previously discussed in the literature. A student's literal understanding of the word 'similar' made her believe that the angles at the vertices could not be the 'same'. The think-alouds provide evidence that this student does not fully understand geometric similarity and perhaps this is due to its treatment in the curriculum at 7th grade and earlier. Both of these characteristics provide insight into ways that the standards such as the CCSS-M might be insufficient guides to student learning and thus further highlight the need for framing instructional planning with LTs.

Level 9 was an exception as it was considered to be a mis fitting level but also that students were generally able to respond to these items using no more than their Level 5 thinking/strategies. Which does imply that the items need to be relocated. Furthermore, some of student thinking from the think-alouds suggests also that Level 9 needs to be moved lower down on the LT. Students can attend to this level of thinking much earlier than anticipated by the existing LT.

### **Summary of Raw Scores from Think Alouds.**

The purpose of the think-aloud interviews was not to derive a simple right or wrong analysis of the students' responses to items. Moreover, the purpose was to discover the extent to which items elicited the thinking and strategies as described in the LT levels. Each colored dot in Figure 45 represents a student's response to an item (blue = full credit, gradient = partial credit, orange = zero credit). The students are organized by their total raw score on the six items that each of them answered during the think-aloud interview. Items were not "equated" across students, so what one student saw as an item from level 5 might not have been equally as difficult as another student's item from the same level. The display offers another way of viewing the data from the think-alouds and triangulates some of the recommendations made the researcher especially in relation to moving Level 9 items to a lower level (Level 5). The display also shows that students with a higher raw total score on the six items tended to get most or all of the questions correct along the LT. However, what is more interesting is the students who had lower raw total score tended to get the questions on the lower part of the LT correct but not so much on the upper part of the LT. The exception was TA2 who got the Level 8 question correct by guessing.

### Level by Think-aloud: An Overview of Student Performance on Items during the Think-alouds



*Figure 45.* Overview of students' raw scores on items they did the think-aloud on.

### Patterns of Item Difficulty Statistics within the *Similarity and Scaling* LT

#### Research Question 2

To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?

Claim: Within the target population, item difficulty increases as the LT-level increases and the item difficulty of multiple items at a given level vary due to instructionally appropriate, construct relevant factors.

This section of the chapter looks at results from the analysis of field-test data collected from all three middle schools over the course of two years. The researcher examines the structure of all the calibrated items ( $n=21$ ) through the lens of empirical item difficulty, where difficulty is

defined by the Rasch model  $\beta$ -parameter. First, however, the choice of IRT model is explained using an output table of global fit indices. Following this is a brief explanation of how the items were scored and the reason for choosing one scoring method over another. Then, the results of the IRT analyses are presented followed by a discussion about why it is imperative to incorporate the recommendations from the think-alouds into the analysis at this stage. Following that, the results from the iterative variable selection method to regression are presented. These results from the regression outline potentially non-conforming items which are subsequently analyzed. Discussion of potentially non-conforming items will include analysis of response patterns from field-testing data. Finally, conclusions from the think-alouds will be integrated with the results from the iterative variable selection method to regression to provide a summary of decisions made in relation to non-conforming items and misplaced levels.

### **Comparison of Fit Statistics.**

Both single group IRT (Rasch) analyses and multi-group (academic years as the groups) analyses were carried out on the data to see whether the data from one academic year to the next needed to be analyzed separately. Model fit was used to select the best model between these two models (Templin, 2018). Table 7 below shows that output from comparing the two models.

Table 7. Comparison of Fit Statistics of Two Confirmatory IRT Models.

<b>Model</b>	<b>AIC</b>	<b>AICc</b>	<b>SBIC</b>	<b>BIC</b>	<b>logLik</b>	<b>2 diff</b>	<b>df</b>	<b>p</b>
<b>1: 1-D Rasch, single group</b>	6720. 06	6730. 33	6770. 67	6923. 03	- 3312.03 2	NA	NA	NA
<b>2: 1-D Rasch, multi group</b>	8244. 53	8268. 03	8319. 39	8544. 75	- 4051.26	- 1478.46	23	1

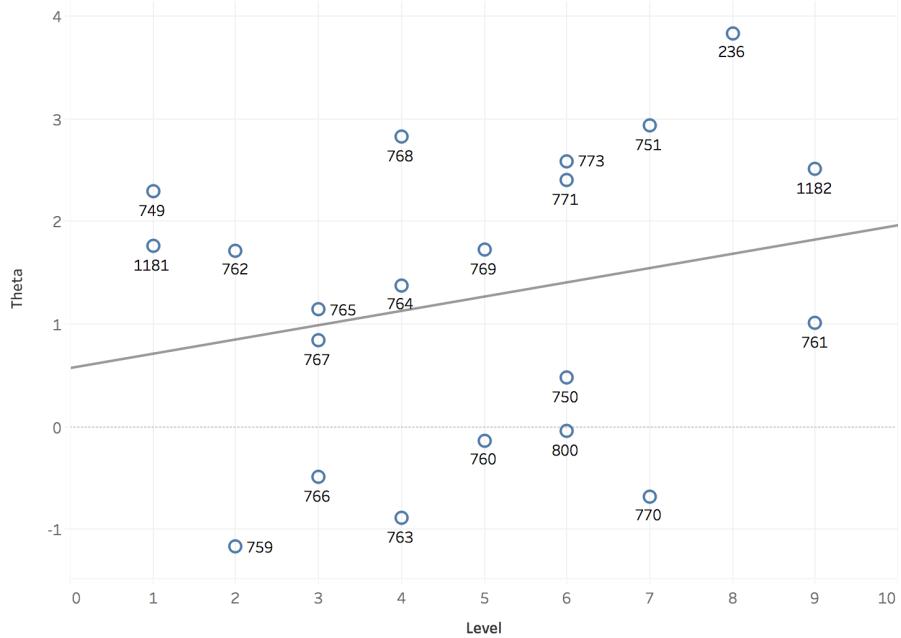
AIC: Akaike's fit index; AICc: corrected AIC; SBIC: sample-size adjusted BIC; BIC: bayesian information criterion; logLik: log likelihood; 2diff: chi-squared difference test; df: degrees of freedom for 2diff; p: probability of a significant difference in fit between models.

de Ayala (1998) recommends that lower AIC implies a better fit. Further, SBIC, BIC, and log likelihood indices are more responsive to differences in sample size. As shown in the table, all of these indices are decisively different. Therefore, the fit statistics favor the 1-D Rasch single group model and the researcher proceeded to use this model to generate item parameters for all 21 items.

### **Results from IRT Rasch model.**

A Rasch model generates parameters for the item difficulty, referred to as b parameters, based on the number of possible points achievable (large b value implies more difficult item). For dichotomous items (no partial credit), the b parameter of an item is the estimate of the theta level (ability parameter) that would be needed to correctly answer the item. For any item that includes two or more parts or is a “select multiple” item type with more than one correct answer choice, the model generates a parameter for getting one out of the two parts correct, and a parameter for getting full credit. Such items are called polytomous items, and for these the b parameter is the estimate of the theta level where the student would have over a 50% chance of getting the that number of points. For this study, the researcher chose to use the b parameter associated with getting all parts of an item correct so that these could be easily compared to the parameters generated for dichotomous items. First, item parameters are presented in Figure 46 to show the distribution of all 21 item parameters by level for the *Similarity and Scaling* LT.

IRT difficulty parameters by LT level for Similarity and Scaling LT (All 21 items)

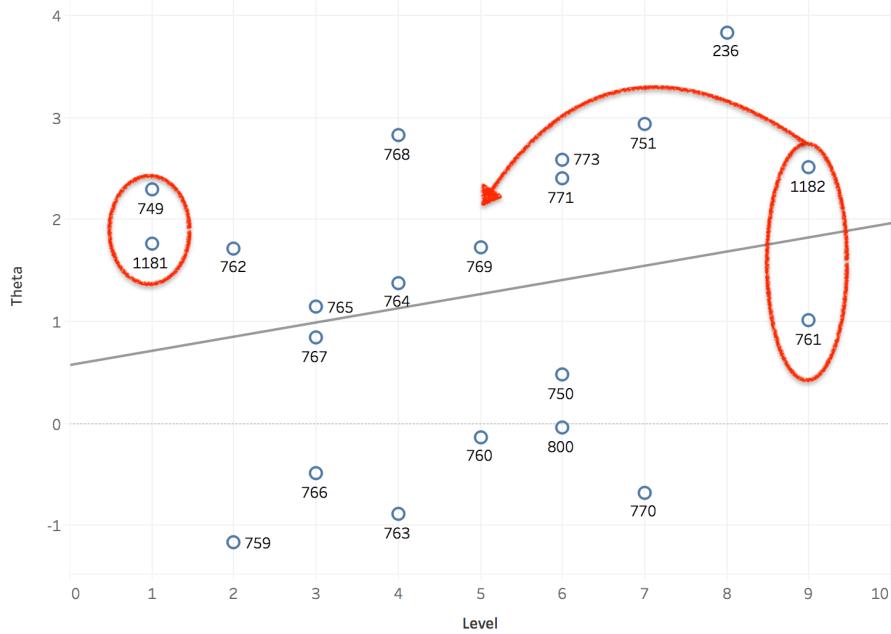


Level vs. Theta. The marks are labeled by Item No.. The data is filtered on Level, which keeps non-Null values only.

Figure 46. IRT (Rasch)  $\theta$  –parameters by LT level for 21 field tested items.

Figure 47 on the other hand, shows an annotated version of the Figure 46. The annotations reference two of the three recommendations made at the end of results section for *Research question 1*. To recap, evidence from the think-aloud interviews suggest that items from Level 1 may have been higher up on the item difficulty scale due to a lack of opportunity to learn. Therefore, the first recommendation was to flag items at Level 1 and continue collecting field test data on them. Evidence from the think-aloud interviews also suggests that the two items written to measure Level 9 were not completely measuring the level and were better aligned to Level 5. Therefore, the items were moved to Level 5. The third recommendation was to re-write items for the relocated Level 9. This last recommendation will require more time and data from field-testing and therefore has not been included in the scope of this study.

IRT difficulty parameters by LT level for Similarity and Scaling LT (All 21 items)

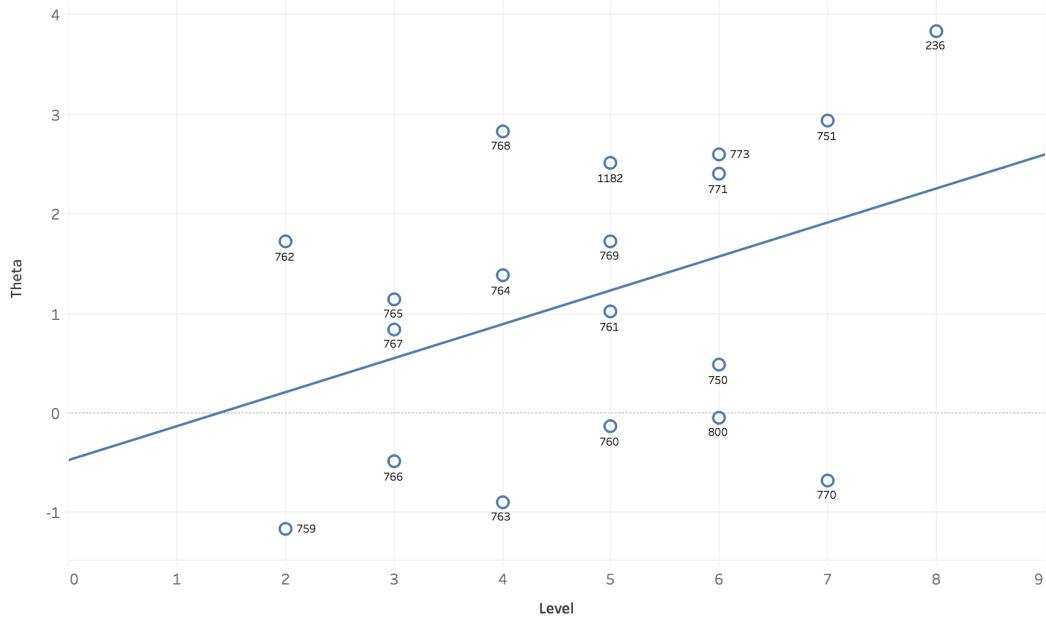


Level vs. Theta. The marks are labeled by Item No.. The data is filtered on Level, which keeps non-Null values only.

Figure 47. IRT (Rasch)  $\theta$  –parameters by LT level for 21 field tested items.

Incorporating two of the recommendations from *Research Question 1* was necessary at this stage in the analysis because it allowed the researcher to build the best model for the initial linear regression (Ratner, 2010). The researcher used a linear regression to examine the relationship between item difficulty ( $\beta$ ) and LT level because a strong positive relationship implies that the pattern of increasing sophistication (content wise) from the first progress level to the  $n^{\text{th}}$  is reflected in the distribution of empirical item difficulties. First, a baseline regression model was run (Figure 48), then the item with the highest absolute residual value was removed, and the regression analysis was repeated. The new  $R^2$ , residual sum of squares (RSS), and slope were then compared to those of the baseline regression model. The ultimate goal was to maximize the  $R^2$ , while also retaining no less than 70% of the items in the pool, so the process of removing and comparing was repeated iteratively until the aforementioned criteria were met.

Revised (baseline) plot of IRT difficulty parameter by LT level for the Similarity and Scaling LT



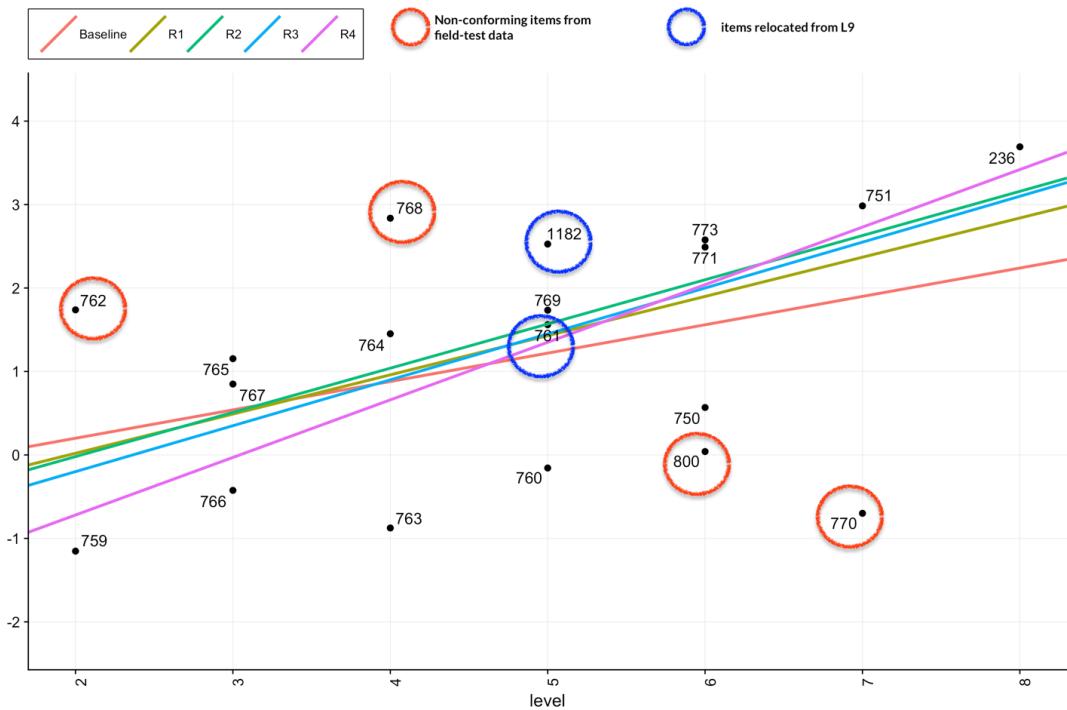
Level1 vs. Theta. Color shows details about ItemID & Item Name 6 (group). The marks are labeled by ItemID. Details are shown for Item Name. The data is filtered on Level1, which keeps non-Null values only. The view is filtered on ItemID & Item Name 6 (group), which excludes (749, ItSM19\_49\_1\_749.d) & (1,181, ItSM19\_49\_1\_1181.d).

Figure 48. The baseline linear regression model applied to the subset of  $19^{10}$  items.

The baseline model generated a moderately positive slope (0.34) and an  $R^2$  value of 0.16, suggesting that the relationship between LT level and item difficulty for this model was somewhat positive but only 16% of the variance in LT level was explained by item difficulty. To improve the fit of the regression line, the absolute residual value for all items was considered; the item with the largest value was removed (Item 770, Level 7,  $n=344$ ). The regression run with the remaining 18 items was an improvement from the baseline model: the slope increased to 0.47, the sum of squares decreased, and the  $R^2$  increased to 0.30. Following the same pattern, three subsequent regressions were run, first removing Item 800 (Level 6,  $n=101$ ), then Item 768 (Level 4,  $n=218$ ), and finally Item 762 (Level 2,  $n=190$ ), with the sum of squares decreasing and  $R^2$  increasing each time, until the  $R^2$  value increased considerably while still retaining close to 80%

<sup>10</sup> Level 1 items have been removed.

of the items (Item 800: 0.38, Item 768: 0.45, Item 762: 0.60). The final regression without Items 770, 800, 768, and 762 displayed marked improvements over the previous model, with slope increasing, sum of squares decreasing,  $R^2$  increasing from 0.16 to 0.60, providing evidence that these items should remain excluded from the final model and be scrutinized to determine whether they should be retained in the item pool. Figure 49 and Table 8 show the results of the sequential regression models generated each time an item was removed.



*Figure 49.* Results of the sequential regression models.

Table 8. Regression Equations, R<sup>2</sup>, and Sum of Squares for iterative regression models.

LT	Model	Slope	Intercept	SS	R <sup>2</sup>
<i>Similarity and Scaling</i>	Baseline (All Items)	0.34	-0.48	32.47	0.16
	R1: Without Item 770	0.47	-0.92	24.63	0.30
	R2: Without Items 770 & 800	0.53	-1.08	20.53	0.38
	R3: Without Items 770, 800, & 768	0.55	-1.30	17.05	0.45
	R4: Without Items 770, 800, 768, & 762	0.69	-2.10	12.36	0.60

Finally, to provide further evidence about the relationship between LT level and item difficulty the Spearman rank correlation coefficient was computed for the final model (R4). The result showed a strong positive correlation: 0.73 ( $p = 0.0018$ ) significant at the  $p < 0.01$ , indicating that item difficulty does increase monotonically along the levels of the LT.

***Discussion of Items from Variable Selection Method to Regression.*** After removing items 770, 800, 768, and 762 sequentially, a regression line that modeled the remaining items in the *Similarity and Scaling* LT provided evidence that concepts became increasingly more sophisticated as one progressed up the levels. The final model (R4) for the *Similarity and Scaling* LT accounted for a 44% increase in the amount of the variance in item difficulty. This improvement from 16% (baseline) suggests that the overall pattern of item difficulty does tend to increase by LT level. In the section below each one of the four potentially non-conforming items is examined more closely. The researcher will consistently analyze each item (770, 800, 768, and 762) by first using the data from the large-scale study, including specific response patterns observed from field-testing. Then the researcher will incorporate the conclusions from the think alouds in making a final decision for each of the items. Information is provided about the current item difficulty parameter of the item paired with what “should” be its item difficulty based on the final regression model (R4).

*Item 770* (Level 7: Uses similarity, proportionality, and scale factors to solve problems involving scale drawings) current  $\beta$ : -2.59,  $\beta$  predicted by regression model: 2.74

This item requires a numeric response, asking students to find the actual distance between two buildings that are 7.2 inches apart on map, given the scale factor 1-inch equals 500 feet.

Over the first two years of field testing, the most common erroneous response (2015-16, 4%,  $n = 5$ ) was 3600 (error with multiplication). In the 2015-16 school year 78% ( $n = 93$ ) of the responses were correct and in the 2016-17 school year 72% ( $n = 101$ ) of the responses were correct. The other harder item at the level required students to apply a fraction scale factor to find the scaled dimensions of a room plan and compare the actual and scaled areas of the room.

Therefore, the researcher recommended that the level's statement be altered to read “*Uses and justifies similarity...*”. In light of this level statement change, the item was recommended for revision because based on students' responses from the think-alouds, students can get this question correct without knowing why their method of multiplication worked. An additional part to the question, asking students to justify their response, would allow for better measurement of student thinking at this level.

*Item 800* (Level 6: Extends similarity relationship from triangles to other shapes, and identifies similar shapes with proportional corresponding sides and congruent corresponding angles) current  $\beta$ : -1.62,  $\beta$  predicted by regression model: 2.05

This question asked students to select two pairs of side lengths, that are in proportion, for a backyard soccer field that was scaled up. There are two correct answer options and two distractors for this item. The two distractors have only one side length enlarged, and the other side is the same dimension as the original soccer field. In school years 2015-16 and 2016-17, response patterns from field test data show that 54% ( $n=56$ ) and 65% ( $n=15$ ) of the students

selected both correct answers choice. The next most frequent response was selecting only one (dimensions that were doubled) out of the two correct answers choices. (2015-16, 29%,  $n=30$  and 2016-17, 17%,  $n=4$ ). The other harder items at the level required students to work with fractional scale factors or identifies similar rectangles with proportional corresponding sides from a larger pool of rectangles. In general, the item appears to be aligned to the level description, but one student think-aloud interview suggests that a student can get this item correct by using a combination of test strategy and a process of elimination. Therefore, an additional, credible distractor would enable this item to better measure student thinking at this level e.g. a pair of side lengths that are larger than the original soccer field but not in proportion. The recommendation is to retire the current item and create another version of it that would have more credible distractors and therefore, ensure that students were using proportional reasoning to solve the problem and not a process of elimination.

*Item 768 (Level 4: Grows or shrinks a similar triangle by doubling, tripling, or halving side lengths and defines similar triangles as having congruent angles and proportional sides) current  $\beta$ : 1.94,  $\beta$  predicted by regression model: 0.66*

The context of this question is a hamster cage that can be re-sized by assembling more and more triangular “building blocks” (base 6in., altitude 4 in.) into proportionally larger hamster cages. The first part requires students to find the base length of a cage that has two triangular building blocks (doubling). The second part requires students to find the altitude of the cage in part one. The third part requires students to find the number of building blocks that would be needed to build a cage that has a 42-inch base. The third part requires students to find the altitude of the cage in part three. Finally, the fifth part asks the total number of triangular building blocks required to build the cage in the third part.

Field test response patterns suggest that about 36% ( $n= 47$ , 2015-16 school year) and 35% ( $n= 15$ , 2016-17 school year) of students correctly answered the first three parts of this item. The next most frequent response pattern is the correct sequence of answers for all five parts (10%,  $n= 13$ , 2015-16 and 8%,  $n= 3$ , 2016-17). The other easier items at this level require students to work with more straightforward doubling of dimensions as described in the LT level. Data from the think-alouds show that students using additive reasoning were only able to get so far with the question and that the fourth part of the question was sufficient for distinguishing students who are beginning to define similar triangles as having congruent angles and proportional sides. The fifth part, on the other hand, contains construct irrelevant thinking as it requires students to use more of functional reasoning (input, output) than simple proportional reasoning which is described at this level of the LT. A recommendation is made to retire the current item and create a new version of it with just the first four parts and not the fifth part.

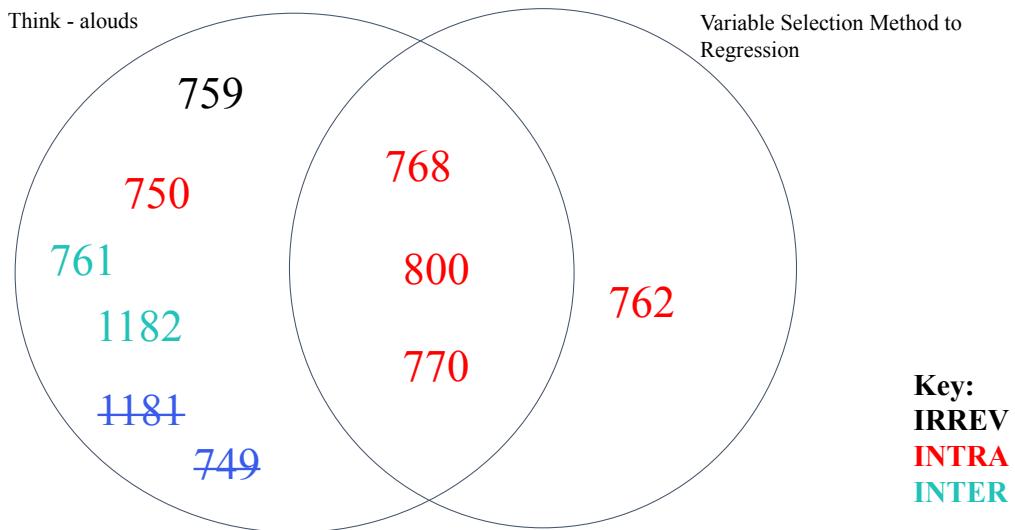
*Item 762 (Level 2: Identifies similar objects as having the “same shape” and justifies by informally scaling, e.g., centering or matching vertices) current  $\beta$ : 1.51,  $\beta$  predicted by regression model: -0.72*

This item requires students to pick one pair of similar triangles from a set of six pairs of triangles (A – F) using visual inspections as no dimensions are provided. There only one correct answer choice (C) but because it is a multiple select item type, students can pick as many answer options as possible. The other easier item at this level requires students to pick pairs of similar shapes but has two pairs of triangles, a pair of circles, and a pair of squares.

Field test data show that 25% ( $n=26$ , 2015-16 school year) and 16% ( $n=10$ , 2017-18 school year) of the responses selected only the correct answer option. Which implies that between 75 and 84% of the responses are selecting more than one option. The most common

response string is a selection of triangles B and C (both pairs are right angle triangles). During the think-alouds, it was noted that this item elicits misconceptions such as “all right-angle triangles are similar” or over-generalization such as “all isosceles triangles are similar.” However, because this is a select multiple item type, it is possible that even though students do pick the correct pair, they are very likely to undo their selection by picking just one more pair of triangles as being similar. There is some way of getting around this without completely retiring the item, especially since it is able to elicit students’ misconceptions pertaining to similar triangles. There may be some redundancy with two of the pairs: D and E and eliminating one of the answer options may make the item less difficult. Additionally, a new pair of triangles that is similar can be introduced to allow for more opportunities to get partial credit for this item.

***Comparison of Recommendations from Think-Alouds and Variable Selection Method to Regression.*** Analysis and recommendations for each of the items based on the think-alouds and recommendations from the variable selection method to regression have been compared for each item in Table 9. Where the item deviated significantly from the level, and the researcher had specifically suggested that the item or parts of it did not measure the level completely, the researcher marked the recommendation with an asterisk (\*).



*Figure 50.* Venn diagram showing the integrated results from two methods.

Based on the think-alouds alone, the researcher would have predicted seven items would be labelled non-conforming and using the variable selection method to regression four items were non-conforming. There overlap of three items showed the value of using multiple methods. In the combined review of items based on both the think-alouds and the review of the non-conforming items from the regression, the researcher was able to resolve and explain the differences. As a result of the two methods, two items were flagged, six items were revised to better address the level, two items were moved to another level, and one level was relocated (Figure 50).

Table 9. Comparison of Recommendations from Think-alouds and Regression.

<b>Item</b>	<b>Level</b>	<b>Classification</b>	<b>Recommendation</b>	<b>Variable Selection Method to Regression</b>
1181	1	FLAG	Collect more data	NA
749	1	FLAG	Collect more data	NA
759	2	IRRE-V	Make changes to pairs of shapes*	-
762	2	ADJUST	Remove a pair of triangles	Revise Item
765	3	FITS	NA	-
766	3	FITS	NA	-
767	3	FITS	NA	-
763	4	ADJUST	Revise the word “distortion”	-
764	4	ADJUST	Improve diagram and question stem clarity	-
768	4	ADJUST	Improve diagram and question stem clarity, Remove last part *	Revise Item
760	5	ADJUST	Add part to check for proportional reasoning	-
769	5	ADJUST	Improve question stem clarity	-
750	6	ADJUST	Change orientations of rectangles*	-
771	6	ADJUST	Improve diagram clarity	-
773	6	FITS	Revise the word “distortion”	-
800	6	ADJUST	Include another credible distractor*	Revise Item
751	7	ADJUST	Change fractional scale factor	-
770	7	FLAG	Add part to check for proportional reasoning*	Revise Item
236	8	FITS	NA	-
241	8	FITS	NA	-
433	8	FITS	NA	-
1182	9	PARTIAL FIT, MISPLACED LEVEL (L5)	Move to Level 5, Rewrite and Relocate Level 9*	-
761	9	PARTIAL FIT, MISPLACED LEVEL (L5)	Move to Level 5, Rewrite and Relocate Level 9*	-

## Distribution of Scores Across Gender and Ethnic Subgroups

### Research Question 3

How do patterns of observed performance differ across students from a diverse sample on LT-based assessments measuring geometric similarity?

**Claim:** The diagnostic assessments that measure students' understanding of geometric similarity are equally sensitive to the construct-relevant student thinking of a diverse sample of students.

The student scores reported in this section come from all three middle schools and have been calculated based on data from two years of field testing the *MM6-8* assessments aligned to the *Similarity and Scaling* LT. A Rasch model has been applied to the data to generate ability parameters<sup>11</sup> ( $\theta$ ) for each of the students based on their performance on the assessment<sup>12</sup>.

Students ability parameters were generated from the entire sample and then results split according to subgroup.

The results to this question have been separated into: a) a comparison of group means i.e. mean theta score between the two genders and b) a comparison of group means between ethnic subgroups. First, a summary of results is provided. These are further supported by histograms of score distributions. Each set of histograms is followed by a report of an analysis of variance (ANOVA) test conducted between the groups means.

---

<sup>11</sup> The mean of the item difficulty parameters ( $\beta$ ) is usually assumed to be the zero for Rasch models (Bond and Fox, 2015). Student ability parameters when plotted on the scale, are read like this: a 50% probability of correctly answering an item located at the same point on the logit scale. For example, a person with an ability parameter of 0 has a 50% probability of succeeding (or failing) on an item with a  $\beta$  parameter of 0.

<sup>12</sup>all test forms that measure the Similarity and Scaling are assumed to be equated. This implies that different groups of students may receive a different set of items on their test but in general the overall of difficulty of the test will be equivalent to other test forms.

**Results: Distribution of scores across gender subgroups.** The results for this section come from 493 students' scores on assessments that contained items from the *Similarity and Scaling LT* ( $n_{males} = 231, n_{females} = 262$ ). The histograms below show distributions of scores for females (Figure 51) and males (Figure 52). The average theta value for females was -0.04 (s.d. = 0.89, min= -2.05, max= 2.03). The average score for males was also -0.04 (s.d.= 0.85, min= -2.63, max= 2.43).

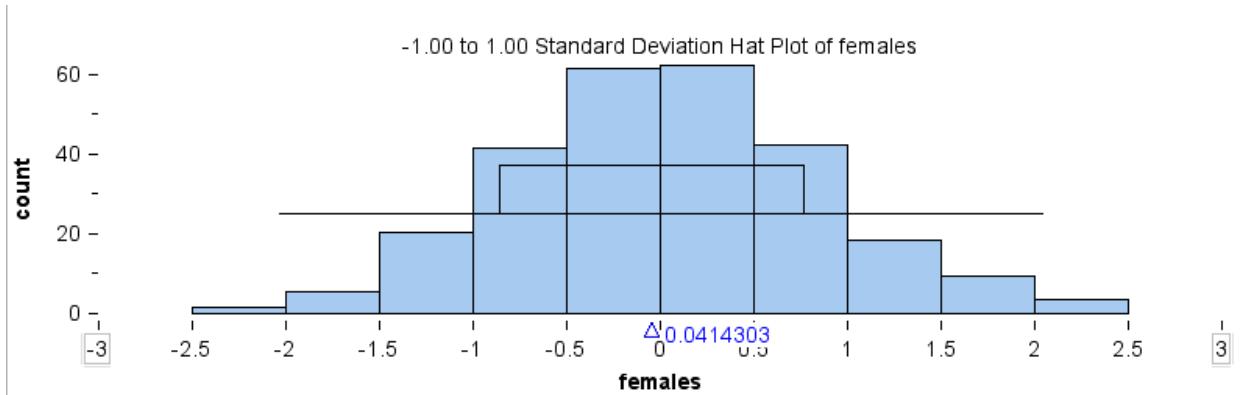


Figure 51. Distribution of theta-scores for female students.

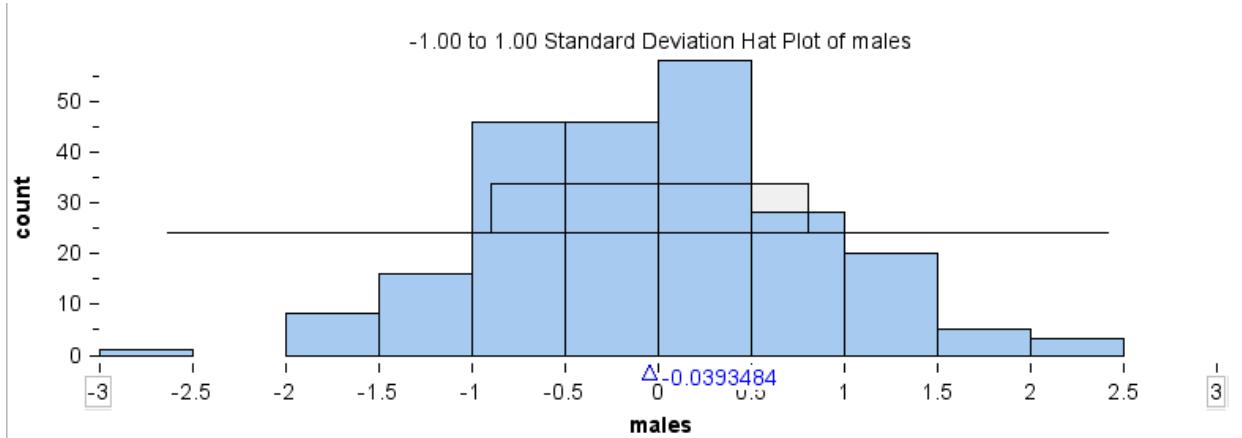


Figure 52. Distribution of theta-scores for male students.

**Discussion.** An ANOVA test was conducted to see whether there is a statistically significant difference between the group means. The output (Table 10) from the statistical package shows that the p-value is 0.282, which implies there is no statistically significant

difference between the means. Post hoc comparisons e.g. Tukey's HSD test are only conducted if results are significant (Oehlert, 2010). Therefore, these results suggest that the *Similarity and Scaling* LT assessments are not biased towards either gender subgroup.

Table 10. Output for ANOVA between gender subgroups.

	SS	df	Mean Square	F value	Pr (>F)
gender	0.8	1	0.8010	1.158	0.282
Residuals	339.7	491	0.6919		

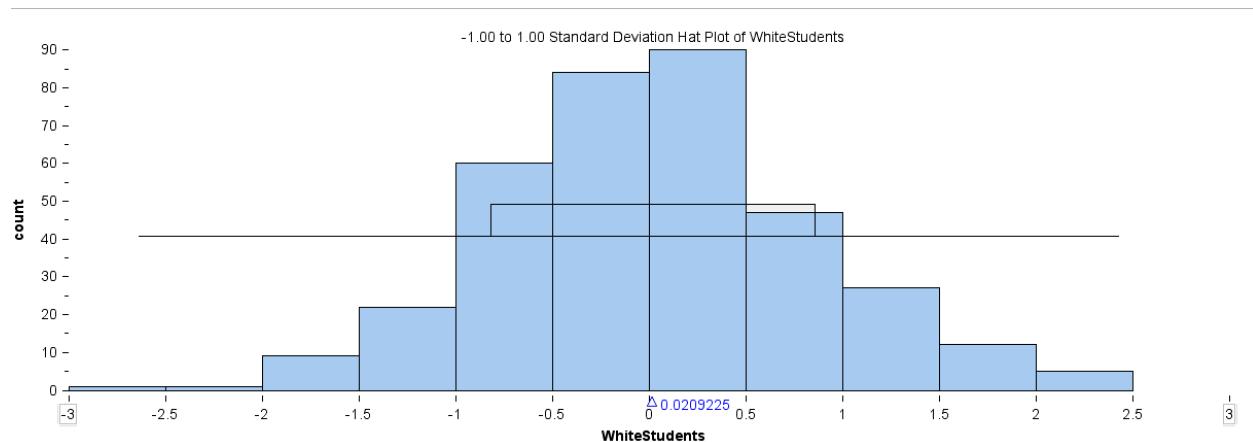
**Results: Distribution of scores across ethnic subgroups.** The same<sup>13</sup> students' scores on the assessments were then analyzed by the following ethnic groups: a) White, b) Hispanic, c) African-American, d) Asian, and e) Mixed race. A summary of these results has been compiled in Table 11. The histograms in Figures 53 - 57 show the distributions for each of these subgroups.

Table 11. Descriptive statistics of student performance on *MM* assessments.

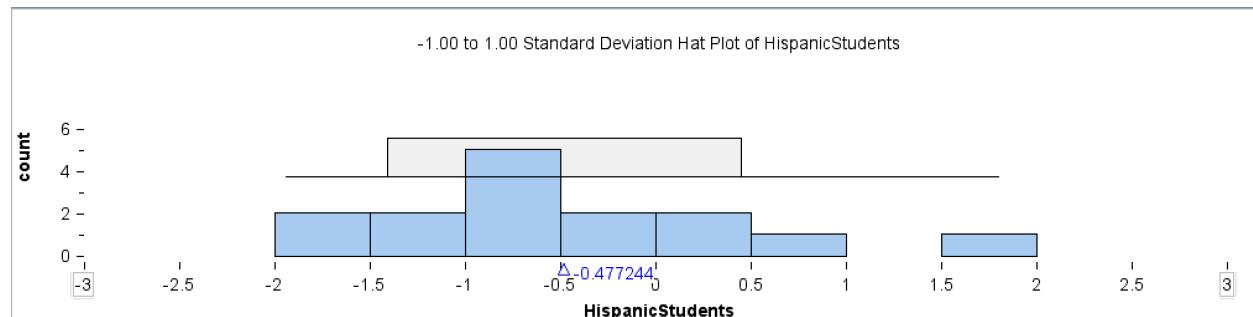
Ethnic Subgroup	n	mean $\theta$	SD	min $\theta$	max $\theta$
White	358	0.02	0.84	-2.63	2.43
Hispanic	15	-0.48	0.92	-1.94	1.81
African-American	54	-0.19	0.81	-1.94	2.03
Asian	27	0.34	0.74	-1.14	1.65
Mixed-race	38	0.08	0.68	-1.49	1.24

---

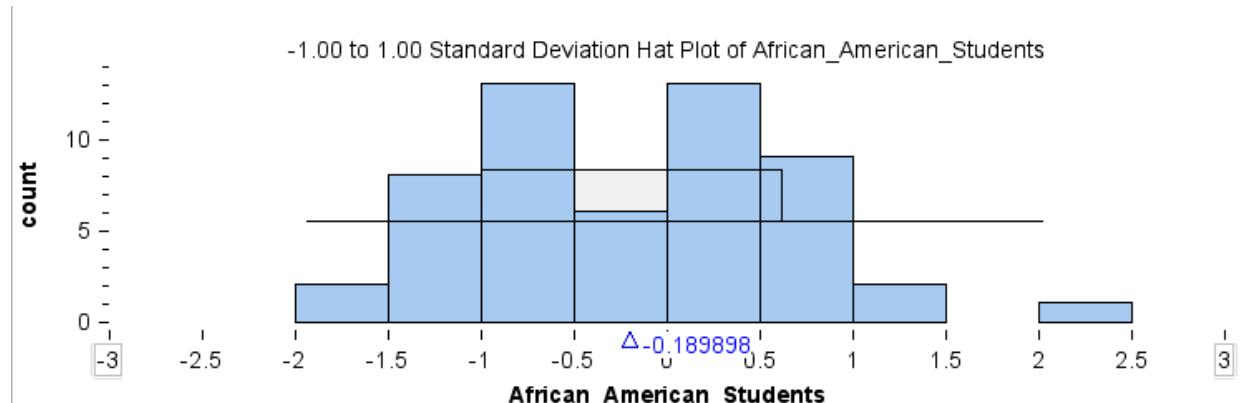
<sup>13</sup> one student of native American ethnic background was omitted from these results so that the researcher was not in breach of privacy regulations



*Figure 53.* Distribution of theta-scores for White students.



*Figure 54.* Distribution of theta-scores for Hispanic students\*.



*Figure 55.* Distribution of theta-scores for African-American students.

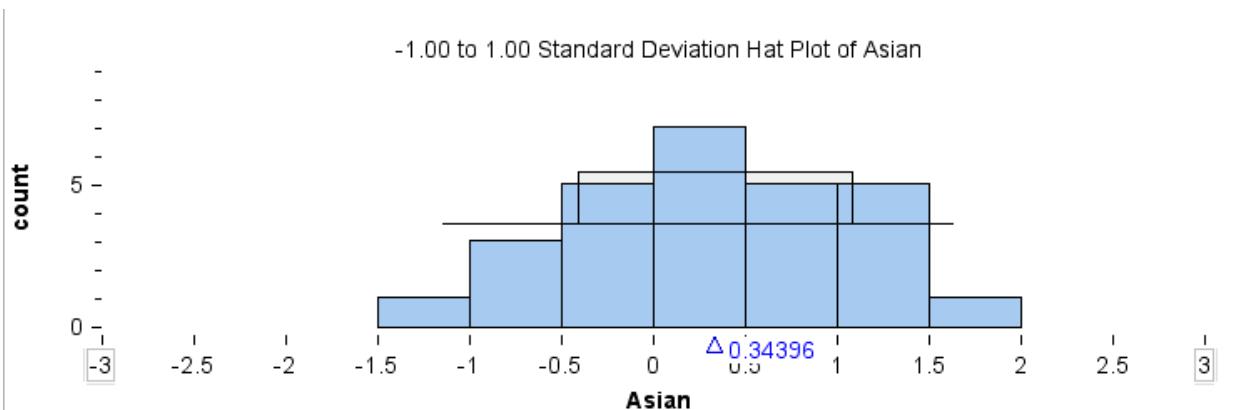


Figure 56. Distribution of theta-scores for Asian students.

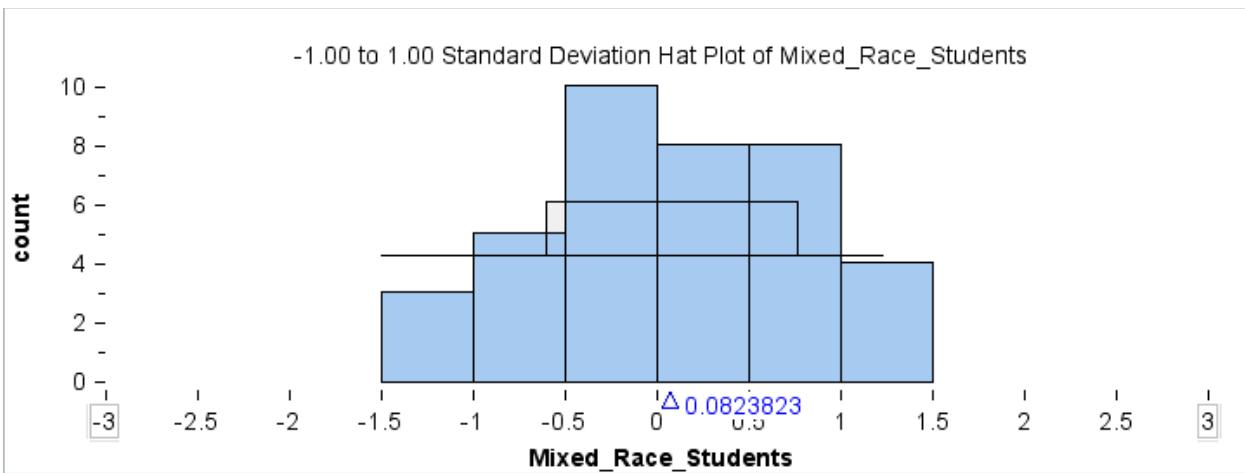


Figure 57. Distribution of theta-scores for Mixed Race students.

**Discussion.** An ANOVA test was conducted to see whether there is a statistically significant difference between the group means. The output (Table 12) from the statistical package shows that the p-value is 0.668, which implies there is no statistically significant difference among the five means. Overall, these results suggest that the *Similarity and Scaling* LT assessment items are not biased towards any specific ethnic subgroup.

Table 12. Output for ANOVA among ethnic subgroups.

	SS	df	Mean Square	F value	Pr (>F)
ethnicity	0.1	1	0.1281	0.185	0.668
Residuals	340.4	491	0.6933		

## **Correlation of Scores from *Similarity and Scaling* Pretest and an Independent Assessment**

### **Research Question 4**

How do the students' results on the *MM 6-8 Similarity and Scaling* assessment compare to those on an independent assessment measuring geometric similarity?

**Claim:** Scores produced by diagnostic assessments that measure LTs are correlated to scores produced by other assessments that measure student understanding in the same domain.

Results for this section come from two assessments that 8th-grade students at District 1 completed. One assessment was a pretest on the *Similarity and Scaling* LT. Note that results from this assessment were also utilized to select a purposeful sample of 18 students for the think-aloud interview reported earlier in this chapter. The other assessment was a set of 15 items which were intentionally selected by the researcher so that they would: a) align to as many levels in *Similarity and Scaling* LT and b) take approximately the same amount of time to complete as the *Similarity and Scaling* pretest. Given that the items were generated by a well-known publisher of assessments for use by districts and schools, the researcher made the assumption that these items were already validated.

Test scores for both assessments were converted to  $\theta$  –scores using a Rasch model. Table 13 below provides a descriptive summary for each assessment while Figure 58 shows the distribution of scores for 131 students who took both assessments.

Table 13. Descriptive Summary of Scores from the pretest and independent assessment.

Assessment	<i>n</i>	mean	SD	min	max
<i>Similarity &amp; Scaling</i> pretest (raw %)	131	32	18	0	82
<i>Similarity &amp; Scaling</i> pretest ( $\theta$ )	131	0.02	0.65	-1.43	1.821
Independent assessment (raw %)	241	44	22	0	93
Independent assessment ( $\theta$ )	131	0.15	0.91	-2.18	1.80

8th grade District 1 Correlation Analysis

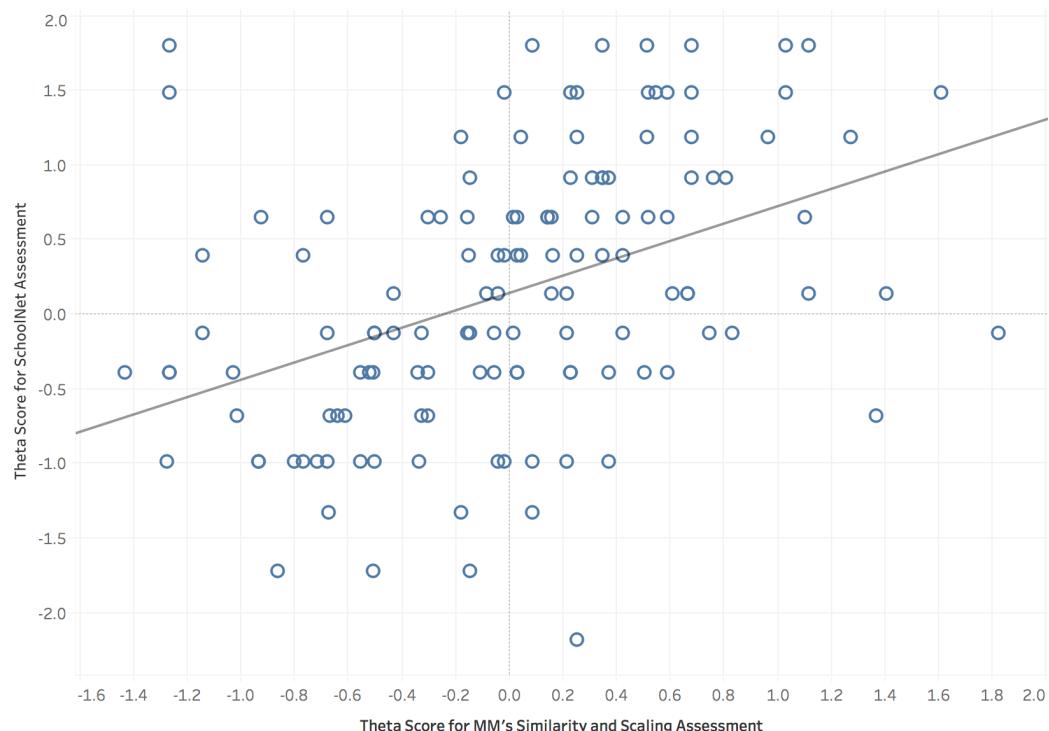


Figure 58.  $\theta$  –scores for pretest vs.  $\theta$  –scores for independent assessment ( $r = 0.42$ ,  $p < .01$ ).

**Discussion.** The results of the correlation indicate that there is a moderate correlation between the scores from both assessments. The result is not overly surprising given that:

the sense in which scores for individual students on different assessments can be said to be comparable to each other or to a fixed standard depends fundamentally on the similarity of the assessment tasks, the conditions of the administration, and their cognitive demands. (Linn, 1993, p. 100)

It is fair to assume that both assessments were taken by students under very similar conditions i.e. both tests were administered during regular mathematics lessons under the supervision of teachers in testing conditions. However, as far as the similarity of the tasks and their cognitive demands were concerned, there were several deviations. For one, the *Similarity and Scaling* pretest spanned all nine levels of the LT whereas the independent assessment item bank had questions that spanned from Level 5 upwards. The reason for this is that the *Similarity and Scaling* LT takes in account students' naive conceptions based on empirical research while the independent assessment item bank consists of items which are written to measure the description of student thinking as articulated by the 7.G.A.1<sup>14</sup> CCSS-M standard.

#### Conclusions and Discussions from Chapter

This chapter was a collection of evidence, established through a variety of data collection tools and procedures of analyses, to support the four validity claims. Each set of results were discussed individually; however, results of the think-alouds were integrated with IRT analyses of the field test data. Results from the think-aloud interviews yielded recommendations for all 21 items in the pool (including two at Level 8 that have not been calibrated). Recommendations from think-alouds then influenced the analysis of data from field testing. A variable selection method to regression was applied to a subset of 19 items and results uncovered four items that were not conforming to the model. These four items were analyzed further from a learning

---

<sup>14</sup> Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale.

sciences perspective and recommendations were made by the researcher to revise all four of the items and collect more response data on them. Further, score distributions were analyzed to check if the *Similarity and Scaling* LT was in any way biased toward certain subgroups of students (gender or ethnic). Finally, 8<sup>th</sup> grade students' scores for an independent assessment measuring 7<sup>th</sup> grade geometric similarity were correlated with their scores for a *Similarity and Scaling* pretest. The overview of these results and what they mean in relation to the overall research question are described in the next chapter.

## **Chapter 5: Discussion and Conclusion**

Learning trajectories (LTs) offer teachers access to an empirical research base on how students learn mathematics, so that they can strengthen their own pedagogical content knowledge (Sarama, Clements, Starkey, Klein, & Wakeley, 2008; Sarama & Clements, 2013; Sarama, Clements, Wolfe, & Spitler, 2016; Wilson, Sztajn, Edgington, Webb, & Myers, 2017; Wilson, Mojica, & Confrey, 2013). Furthermore, teachers can become more student-centered in their instructional practices by organizing their curriculum in a manner that values different approaches and provides all students the appropriate opportunities to gain competency in their understanding of big ideas in mathematics (Nguyen & Confrey, 2014; Olson, 2014).

Simply providing teachers access to LTs is not sufficient in effecting the change that is needed to improve learning opportunities for students. One way of effecting change may be to provide teachers with a set of assessments aligned to a theory of learning (LTs) as a means to drive their instructional decision making with data about student learning (National Research Council, 2003; Pellegrino, Chudowsky & Glaser, 2001). The fundamental premise is that these assessments provide teachers and students with reliable and valid feedback about gaps in learning, thus paving the way for more targeted modifications in instruction based on timely and accurate data.

In order for LT-aligned assessments to support teachers in being more learner-centered, they must be validated. However, researchers call for re-conceptualizing validity in the context of classroom assessments, rather than simply importing the notions of validity and reliability that have been conceptualized in the context of large-scale assessments (Brookhart, 2003; 2018; Pellegrino, DiBello & Goldman, 2016).

This study was situated in a larger program of ongoing validation studies (Confrey, under review) that investigate claims associated with the design and use of a digital learning system (DLS) *MM 6-8*. The SUDDS team's approach has been to identify one or more purposes of *MM 6-8* and then address the associated components in the validation framework: cognition, instruction/implementation, and inference (adapted from Pellegrino et al., 2016). For example, one purpose of *MM 6-8* is to inform teachers of class progress by LT. In the context of this purpose, the specific objective of this study was to investigate the validity of an assessment of an LT developed for students in the middle grades who are studying geometric similarity. The overall research question investigated was:

*To what extent do the assessments aligned to the Similarity and Scaling learning trajectory within the digital learning system *MM 6-8* provide a user with valid information about student thinking along a continuum of learning?*

Four research questions were investigated, and these were associated with four distinct but related validity claims:

1. *To what extent do data collected from think-aloud interviews correspond to the descriptions of students' understanding of geometric similarity as defined in the LT levels in *MM 6-8*?*

**Claim:** The items in the diagnostic assessments that measure students' understanding of geometric similarity are aligned to the corresponding LT levels as defined in the DLS *MM 6-8*.

2. *To what extent do the data collected from the ongoing field testing support the structure of the LT (within and between progress levels) on geometric similarity?*

**Claim:** Within the target population, item difficulty increases as the LT level increases and the item difficulty of multiple items at a given level vary due to instructionally appropriate, construct-relevant factors.

3. *How do patterns of observed performance differ across students from a diverse sample on LT-based assessments measuring geometric similarity?*

**Claim:** The diagnostic assessments that measure students' understanding of geometric similarity are equally sensitive to the construct-relevant student thinking of a diverse sample of students.

4. *How do the students' results on the MM 6-8 Similarity and Scaling assessment compare to those on an independent assessment measuring geometric similarity?*

**Claim:** Scores produced by diagnostic assessments that measure LTs are correlated to scores produced by other assessments that measure student understanding in the same domain.

The findings of this study have implications for teachers, learning scientists, classroom assessment developers, measurement theorists, and writers of standards. This chapter will present a brief overview of the study, discuss key findings, discuss implications of the findings for various audiences, detail limitations of this study, offer recommendations for future research, and close with final remarks on the significance of such research. Implications integrate findings from two sub-studies: the investigation of student thinking about geometric similarity based on the think-alouds and the investigation of the claims about the structure of the Similarity and Scaling LT.

## **Overview of Study**

This study analyzed qualitative data from think-aloud interviews with 8<sup>th</sup>-grade students from a research partner school to examine the first claim about the assessment items being aligned to the corresponding levels of the *Similarity and Scaling* LT. These data were complemented with a quantitative IRT analyses of the same items, based on two years of field test-data to examine the second claim about how the items support the structure of the LT (within and between progress levels) on geometric similarity. Additionally, quantitative data from field testing were analyzed to examine the claim that the assessments aligned to the *Similarity and Scaling* LT were equally sensitive to the construct-relevant student thinking of a diverse sample of students. Lastly, scores produced by the *Similarity and Scaling* LT pretest taken by 8<sup>th</sup>-grade students at one middle school were correlated to scores produced by an independent assessment that measured student understanding in the same domain to examine the final claim.

### **Student learning of middle grades geometric similarity**

The SUDDS team has developed this LT based on empirical research on student learning (Chazan, 1988; Confrey, 1992; Lehrer et al. 2002; turnonccmath, 2018). Overall, data from the think alouds confirmed, for the most part, that the items corresponded to the levels. Items written to measure Level 9 were an exception as they did not completely capture the full extent of the LT level (Level 9: *Uses informal arguments about angles to establish the AA criteria for similarity and solves similar triangle problems in context*). However, there was another issue with the placement of Level 9. That is, in order to conform with the grade level mapping of the standards, Level 9 was placed at the top of the LT. However, data suggest that it should have been placed immediately after level 5 (*Defines similar triangles as having equal angles and sides in proportion and, given one pair of corresponding sides, solves for unknown corresponding sides*).

The authors of the standards delayed the uses of informal arguments using the AA criteria inordinately by placing the standard in eighth grade. Perhaps, their decision was made in an attempt to combine AA arguments into one single standard with angle sum of triangles, exterior angle, and transversals (*8.G.A.5 Use informal arguments to establish facts about the angle sum and exterior angle of triangles, about the angles created when parallel lines are cut by a transversal, and the angle-angle criterion for similarity of triangles.*)

Another insight concerns the foundational learning needed for a robust understanding of geometric similarity i.e. geometric transformations are also not covered until 8th grade (*8.G.A.2: Understand that a two-dimensional figure is congruent to another if the second can be obtained from the first by a sequence of rotations, reflections, and translations; given two congruent figures, describe a sequence that exhibits the congruence between them*). In MM 6-8, the *Similarity and Scaling* LT and a related LT, *Transformations of Geometric Shapes* are located within the same relational learning cluster. The rationale for placing these LTs side-by-side comes from more empirical research in student learning of geometric similarity, which suggests that the study of transformations of shapes is foundational to the formal study of similarity (Usiskin, 1972; Seago et al., 2013). However, as long as the standards continue to separate these two inter-linked ideas (similarity and transformations) across two different grades, 7<sup>th</sup> grade students will not have the opportunity to leverage the underpinnings of similarity as found in study of geometric transformations.

The think-aloud study and its analysis did reveal two other insights into student understanding of geometric similarity (and its related concept of congruency) that have not been reported in the literature reviewed for this study. The first insight is that some 8th-grade students incorrectly believe that if two shapes have the same area, then they must be congruent. This

result became evident based on an item analysis of items aligned to Level 1 where it was offered as an option and two students chose it. Furthermore, in the interviews, two students expressed their strong attraction to this belief which may be due to a misapprehension that while congruence ensures equal area, equal area does not ensure congruence. This misconception can go undetected in the middle grades if teachers ground their instructional planning on a literal<sup>15</sup> interpretation of the CCSS-M standard associated with the LT (*Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale*). A literal treatment of the standards can also lead to students having a procedural understanding of why proportionality statements help them solve for missing sides. Moreover, such a procedural understanding of geometric similarity lends itself to typical pitfalls often articulated by teachers, e.g. the student “forgot” to work with corresponding sides of shapes when setting up a proportionality statement or the student used additive reasoning instead of multiplicative reasoning when enlarging or shrinking shapes (Son, 2013).

The second insight came from a student’s literal understanding of the word ‘similar’ made her believe that the angles at the vertices could not be the ‘same’. The earlier levels that include the visual inspection of shapes to check for similarity provide the foundations for what it means for two shapes to be similar. Students, given the opportunity, can recognize that the numbers marking each of the vertices are not arbitrary labels but that they are measures that need to remain invariant across two or more shapes in order for them to be classified as “similar”.

Both of these, students’ misconception about area as a sufficient condition of congruences and the student’s literal understanding of the word “similar”, provide insight into

---

<sup>15</sup> The dictionary meaning of the word literal is used i.e. following the words of the original very closely and exactly

ways that the standards might be insufficient guides to student learning and thus demonstrate the need for the LTs to inform instruction.

Lastly, analyses of the regression of item difficulties against LT levels raised a critical issue in so far as the relationship among three different constructs: proportionality, similarity, and scale diagrams. The interplay of these constructs within specific items (e.g. Level 7 items about scale drawings) indicate that such items may be measuring constructs other than students' understanding of similarity and scale. The think-alouds conducted for the Level 7 items did not reveal whether students were using 6<sup>th</sup> grade ratio reasoning e.g. building up of ratios or 7<sup>th</sup> proportional reasoning around the constant of proportionality as a scale factor. However, it did reveal that students are able to answer scale drawing questions without invoking any substantial proportional reasoning. As such there are two key avenues that need to be pursued. The first, as per the recommendation from the think-aloud interviews, is to revise any questions related to scale drawings so that students are required to demonstrate proportional reasoning (either 6<sup>th</sup> or 7<sup>th</sup> grade) to respond to questions. The second avenue involves a better understanding of the relationship among these three constructs. For this, clinical interviews or design/teaching experiments would have to be employed.

### **Validating learning trajectories-based assessments**

Based on the think-alouds alone, the researcher would have predicted 7 items would be labelled non-conforming and using the variable selection method to regression 4 items were non-conforming. The overlap of items between the two methods showed the value of using multiple methods to identify items that perhaps are not measuring the given level. In the combined review of items based on both the think-alouds and the review of the non-conforming items from the regression, the researcher was able to resolve and explain the differences. As a result of the two

methods, 2 items were flagged, 6 items were revised to better address the level, 2 items were moved to another level, and one level was relocated.

Results from the analyses of these think-alouds, when integrated with quantitative IRT analyses, provided a powerful way to view the items as not just parameters on a scale but as key indicators of the breadth (Level 1 to Level 9 sophistication) as well as depth (flexibility to work with a range of construct relevant items at a given level) of student understanding of a concept. Indeed, what this dual method of integrating qualitative learning science analyses with quantitative IRT analyses allows is a way to ensure that an ill-fitting item is deemed so by characteristics defined by both learning sciences perspective and a measurement perspective. In the case of learning trajectories such as *Similarity and Scaling*, which are constructed from a review of the empirical learning science research, it is imperative that an interdisciplinary approach to building measures is taken. That is, methods and strengths of each discipline must be combined to meet the needs of this emerging genre of assessment: classroom assessment. The study and its findings are beneficial for the field of mathematics education specifically because it offers a method for developing and validating an LT through existing literature and capitalizing on measurement theory.

In terms of Claim 4, the study attempted to link scores from two assessments that were not designed for the same purposes (Lim, 1993; Mislevy, 1992). However, it should be noted that the contemporary nature of this research means that there was no other comparable, LT-based assessment to work with. In fact, one conclusion that could be drawn from the poor correlation between A and B is that the *Similarity and Scaling* LT assessments are distinct from assessments that are not based on a theory of learning.

The study also highlights the importance of incorporating, in due course, other sources of validity evidence. For example, it will be necessary to demonstrate that both teachers and students are able to draw meaningful conclusions from scores generated by LT-aligned assessments. Such evidence can include classrooms observations when teachers are discussing the assessment reports and interviews of both teachers and students analyzing their class/student reports. It is also vital that both these key stakeholders act on their understanding of the reports to make decisions about their next steps. Classroom observations about how teachers address gaps in student understanding and student interviews about what they intend to do individually will be relevant sources of validity evidence. Ultimately, evidence about student growth will be needed to demonstrate that the use of such LT-aligned, classroom-based assessments are indeed helping students gain a more sophisticated understanding of a given mathematical concept. Studies of pre- and posttest data from assessments will be critical evidence of such growth.

### **Limitations**

The researcher made a sincere attempt to ensure that the limitations of this study were minimal, however, no study is free of any limitations (Punch, 2000). First, the generalizability of these results is limited by the sample size of items. There were only 21 items aligned to a 9-level LT and some levels (e.g., Level 5) had more items aligned to them than others (e.g., Level 8). Furthermore, the data collected for each of the items was also not balanced across items, i.e., some items had more observed responses than others ( $\text{min}= 113$ ,  $\text{max}=344$ ). Qualitative data from think-aloud interviews were coded and analyzed by the researcher in isolation and thus interpretations of student thinking are limited due to the researcher's potential biases. In addition, only 18 students were interviewed from one school, limiting the generalizability of claims supported by think-alouds. Future research should aim to replicate this study with more items

that are equally spread along the LT. Think-aloud transcripts should be coded by multiple coders to minimize bias in the interpretations of student learning.

### **Closing Remarks**

We live in a world where data-driven decision making is omnipresent, so much so that we have companies whose sole business model depends on buying and selling user data. I will not comment on the ills of these models but reckon with the fact that technological advances in the field of big data analysis have their upsides and downsides. Really, big data is a powerful tool that could transform education in ways good and bad. Moreover, it's inevitable (Lohr, 2012). So, the questions are not: a) whether to use big data in classrooms, but given that it's already happening, how can learning scientists and measurement theorists join forces so that data is used responsibly to serve the best interests of our students, and b) how can we enable teachers to be data literate, but can innovations in classroom assessments help teachers' develop habits of mind with *a relentless focus on evidence* of student learning (Bocala & Parker Boudet, 2015, p. 7.)?

## References

- Arieli-Attali, M., Wylie, E. C., & Bauer, M. I. (2012). (2012). The use of three learning progressions in supporting formative assessment in middle school mathematics. Paper presented at the *Annual Meeting of the American Educational Research Association*, Vancouver, Canada.
- Battista, M. T. (1999). Fifth graders' enumeration of cubes in 3D arrays: Conceptual progress in an inquiry-based classroom. *Journal for Research in Mathematics Education*, 30(4), 417-448. 10.2307/749708 Retrieved from <http://www.jstor.org.prox.lib.ncsu.edu/stable/749708>
- Battista, M. T. (2004). Applying cognition-based assessment to elementary school students' development of understanding of area and volume measurement. *Mathematical Thinking and Learning*, 6(2), 185-204.
- Battista, M. T. (2011). Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. *The Mathematics Enthusiast*, 8(3), 507-570.
- Battista, M. T., & Clements, D. H. (1996). Students' understanding of three-dimensional rectangular arrays of cubes. *Journal for Research in Mathematics Education*, 27(3), 258-292. 10.2307/749365 Retrieved from <http://www.jstor.org/stable/749365>
- Battista, M. T., Clements, D. H., Arnoff, J., Battista, K., & Caroline Van, A. B. (1998). Students' spatial structuring of 2D arrays of squares. *Journal for Research in Mathematics Education*, 29(5), 503-532. 10.2307/749731 Retrieved from <http://www.jstor.org.prox.lib.ncsu.edu/stable/749731>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. 10.1080/0969595980050102 Retrieved from <http://dx.doi.org/10.1080/0969595980050102>

- Bocala, C., & Boudett, K. P. (2015). Teaching educators habits of mind for using data wisely. *Teachers College Record*, 117(4), 1-20.
- Bond, T., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.
- Branch, J. L. (2000). *Investigating the information-seeking processes of adolescents*//dx.doi.org.prox.lib.ncsu.edu/
- Brookhart, S. M. (2015). *How to make decisions with different kinds of student assessment data*: ASCD.
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5-12. 10.1111/j.1745-3992.2003.tb00139.x Retrieved from http://dx.doi.org.prox.lib.ncsu.edu/10.1111/j.1745-3992.2003.tb00139.x
- Cai, J. (Ed.). (2017). *Compendium for research in mathematics education*. Reston, Virginia: NCTM.
- Carney, M. B., & Smith, E. Using instrument development processes to iteratively improve construct maps: An example in proportional reasoning.
- Chazan, D. (1988). (1988). Proof and measurement: An unexpected misconception. Paper presented at the *Twelfth Conference of the International Group for the Psychology of Mathematics Education*, Veszprem, Hungary. 224.
- Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (1st ed., pp. 420-464). Reston, Virginia: NCTM.

- Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, 6(2), 81-89.
- Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the building blocks project. *Journal for Research in Mathematics Education*, 38(2), 136-163. Retrieved from <http://www.jstor.org/stable/30034954>
- Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach* (2nd ed.). New York: NY: Routledge.
- Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, 42(2), 127-166.
- Common Core Standards Writing Team. (2016). *Progressions for the common core state* (). Tucson, AZ: Institute for Mathematics and Education. Retrieved from [http://commoncoretools.me/wp-content/uploads/2016/03/ccss\\_progression\\_g\\_7\\_hs\\_2016\\_03\\_27.pdf](http://commoncoretools.me/wp-content/uploads/2016/03/ccss_progression_g_7_hs_2016_03_27.pdf)
- standards in mathematics (draft). grades 7–8, HS, geometry.
- Confrey, J., Maloney, A. P., & Nguyen, K. H. (2014). Learning trajectories in mathematics. In A. P. Maloney, J. Confrey & K. H. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. xi–xxi). Charlotte, NC: Information Age Publishing.
- Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education*, 16, 3-56. Retrieved from <http://www.jstor.org/stable/1167350>
- Confrey, J. (1992). First graders' understanding of similarity. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco, CA.

Confrey, J. (2006). The evolution of design studies as methodology. In K. R. Sawyer (Ed.), (pp. 135). Cambridge: Cambridge University Press.

Confrey, J. (2015). Some possible implications of data-intensive research in education—The value of learning maps and evidence-centered design of assessment to educational data mining. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 79-87). Washington, DC: Computing Research Association.

Confrey, J., & Carrejo, D. (2005). Chapter 4: Ratio and fraction: The difference between epistemological complementarity and conflict. *Journal for Research in Mathematics Education Monograph, 13* Retrieved from <http://www.jstor.org/stable/30037732>

Confrey, J., Gianopoulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *Zdm, , 1-18.* 10.1007/s11858-017-0869-1

Confrey, J., Hasse, E., Maloney, A. P., Nguyen, K. H., & Varela, S. (2010). (2010). Executive summary: Recommendations from the conference. Paper presented at the *Designing Technology-Enabled Diagnostic Assessments for K-12 Mathematics*, North Carolina State University.

Confrey, J., Maloney, A. P., Wilson, P. H., & Nguyen, K. H. (2010). (2010). Understanding over time: The cognitive underpinnings of learning trajectories. Paper presented at the *Annual Meeting of the American Education Research Association*, Denver, CO.

Confrey, J., & Maloney, A. (2012). Next generation digital classroom assessment based on learning trajectories. In C. Dede, & J. Richards (Eds.), *Steps toward a digital teaching platform* (pp. 134-152). New York: NY: Teachers College Press.

- Confrey, J., & Maloney, A. (2017). (2017). Researching and using learning progressions/trajectories. Paper presented at the *PME 41 Research Forum*, Singapore.
- Confrey, J., Maloney, A. P., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *Zdm*, 46(5), 719-733.
- Confrey, J., Maloney, A., Nguyen, K., Mojica, G., & Myers, M. (2009). (2009). Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories. Paper presented at the *In 33rd Conference of the International Group for the Psychology of Mathematics Education*, Thessaloniki, Greece. 345-352.
- Confrey, J., & Toutkoushian, E. A validation approach to middle-grades learning trajectories within a digital learning system applied to the “Measurement of characteristics of circles” .
- Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), 3-21.
- Corcoran, Mosher, & Rogat. (2009). *Learning progressions in science: An evidence-based approach to reform*. (Research Report No. 63). Philadelphia, PA: Consortium for Policy Research in Education. 10.12698/cpre.2009.rr63
- Cox, D. C. (2013). Similarity in middle school mathematics: At the crossroads of geometry and number. *Mathematical Thinking and Learning*, 15(1), 3-23. 10.1080/10986065.2013.738377  
Retrieved from <http://dx.doi.org/10.1080/10986065.2013.738377>
- Cox, D. C., & Lo, J. (2014). Detecting distortion: Bridging visual and quantitative reasoning on similarity tasks. *Mathematics Education Research Journal*, 26(1), 1-23.
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (Fourth ed.) SAGE Publications.

- Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction.* ().CPRE Research Reports.
- de Ayala, R. J. (2009). *Methodology in the social sciences. the theory and practice of item response theory.* New York: NY: Guilford Press.
- Dede, C. (2014). *The role of digital technologies in deeper learning.* (Students at the Center: Deeper Learning Research Series). Boston: MA: Jobs for the Future.
- Dienes, Z. P. (2001). 6 stages with rational numbers. *Mathematics in School*, 30(1), 41-45.  
Retrieved from <http://www.jstor.org/stable/30212135>
- Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development.* Lawrence Erlbaum Associates Publishers.
- Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, 47(2), 123-182.  
10.1080/03057267.2011.604476 Retrieved from  
<http://dx.doi.org.prox.lib.ncsu.edu/10.1080/03057267.2011.604476>
- Ellis, A. B., Ozgur, Z., Kulow, T., Dogan, M. F., & Amidon, J. (2016). An exponential growth learning trajectory: Students' emerging understanding of exponential growth through covariation. *Mathematical Thinking and Learning*, 18(3), 151-181.
- English, L. D. (Ed.). (2008). *Handbook of international research in mathematics education* (2nd ed.). New York: Routledge.
- English, L. D., & Kirshner, D. (Eds.). (2015). *Handbook of international research in mathematics education* (3rd ed.). New York: Routledge.

- Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 223-241). New York, NY: Cambridge University Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis* (2nd ed.). Cambridge, MA: MIT Press.
- French, D. (2004). *Teaching and learning geometry: Issues and methods in mathematics education*. London: Continuum.
- Grouws, D. A. (Ed.). (1992). *Handbook of research on mathematics teaching and learning* (1st ed.). New York: Macmillan.
- Guion, R. M. (1977). Content Validity—The source of my discontent. *Applied Psychological Measurement*, 1(1), 1-10. 10.1177/014662167700100103
- Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, 89(2), 140.
- Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. *Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO)*,
- Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* (). Washington, DC: ERIC. Retrieved from <https://eric.ed.gov/?id=ED543063>
- Hess, K., & Kearns, J. (2010). *An introduction to the learning progressions frameworks designed for use with the common core state standards in mathematics K-12*. (). Dover, NH: National Alternate Assessment Center at the University of Kentucky and the National Center for the

- Improvement of Educational Assessment. Retrieved from  
[http://www.nciea.org/sites/default/files/publications/Math\\_LPF\\_KH11.pdf](http://www.nciea.org/sites/default/files/publications/Math_LPF_KH11.pdf)
- Heubert, J. P., & Hauser, R. M. (1999). *High stakes testing for tracking, promotion, and graduation* (). Washington, DC: National Academy Press.
- Izsák, A., & Templin, J. (2016). Coordinating conceptualizations of mathematical knowledge with psychometric models: Opportunities and challenges. In A. Izsák, J. Remillard & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (). Reston, VA: NCTM.
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112-133.  
10.1177/1558689806298224 Retrieved from [http://dx.doi.org/10.1177/1558689806298224](http://dx.doi.org/10.1177/10.1177/1558689806298224)
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135-170.  
10.1207/s15366359mea0203\_1 Retrieved from  
[http://dx.doi.org/10.1207/s15366359mea0203\\_1](http://dx.doi.org/10.1207/s15366359mea0203_1)
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. 10.1111/jedm.12000
- Lai, E. R., Kobrin, J. L., DiCerbo, K. E., & Holland, L. R. (2017). Tracing the assessment triangle with learning progression-aligned assessments in mathematics. *Measurement: Interdisciplinary Research and Perspectives*, 15(3-4), 143-162.  
10.1080/15366367.2017.1388113
- Larson, M. (2016, October 25). Bringing needed coherence and focus to high school mathematics; Retrieved from <https://www.nctm.org/News-and-Calendar/Messages-from-the->

President/Archive/Matt-Larson/Bringing-Needed-Coherence-and-Focus-to-High-School-Mathematics/

- Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education*, 99(3), 432-437.
- Lehrer, R., Jenkins, M., & Osana, H. (1998). Longitudinal study of children's reasoning about space and geometry. In R. Lehrer, & D. Chazan (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 137-167). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lehrer, R., Kim, M. J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey & K. H. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (1st ed., pp. 31-59). Charlotte, NC: Information Age Publishing.
- Lehrer, R., Strom, D., & Confrey, J. (2002). Grounding metaphors and inscriptional resonance: Children's emerging understanding of mathematical similarity. *Cognition and Instruction*, 20(3), 359-398. 10.1207/S1532690XCI2003\_3
- Lester, F. K. (Ed.). (2007). *Second handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics* (2nd ed.). Charlotte, NC: Information Age Publishing.
- Lo, J., Cox, D. C., & Mingus, T. (2006). A conceptual-based curricular analysis of the concept of similarity. Paper presented at the *Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Yucatan, Mexico.

- Lobato, J., & Walters, C. D. (2017). A taxonomy of approaches to learning trajectories and progressions. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 74-101). Reston, VA: National Council of Teachers of Mathematics.
- Lohr, S. (2012, February 11.). The age of big data. *New York Times*
- Martinková, P., Drabinová, A., Liaw, Y., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, 16(2), rm2.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: NY: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. 10.1037/0003-066X.50.9.741
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement*, 1(1), 3-62.
- National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment - workshop report*. Washington, DC: The National Academies Press. 10.17226/10802
- Nichols, P. D., Kobrin, J. L., Lai, E., & Koepfler, J. (2016). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp, & J. P. Leighton (Eds.), *The wiley handbook of cognition and assessment* (pp. 13-40) John Wiley & Sons.
- Panorkou, N., Maloney, A. P., Confrey, J., & Platt, D. (2014). (2014). Developing elementary students' reasoning of geometric transformations through dynamic animation. Paper

presented at the *Proceedings of the 3rd International Constructionism Conference*, Vienna, Austria.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.10.17226/10019 Retrieved from  
<https://www.nap.edu/catalog/10019/knowing-what-students-know-the-science-and-design-of-educational>

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, 51(1), 59-81. 10.1080/00461520.2016.1145550 Retrieved from  
<http://dx.doi.org/10.1080/00461520.2016.1145550>

Pham, D., Bauer, M., Wylie, C., & Wells, C. Using cognitive diagnosis models to evaluate a learning progression theory&nbsp;

Pothier, Y., & Sawada, D. (1990). Partitioning: An approach to fractions. *The Arithmetic Teacher*, 38(4), 12-16. Retrieved from <http://www.jstor.org/stable/41195032>

Punch, K. (2000). *Developing effective research proposals* Sage.

Ratner, B. (2010). Variable selection methods in regression: Ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing*, 18(1), 65-75. 10.1057/jt.2009.26 Retrieved from <https://doi.org/10.1057/jt.2009.26>

Richard A. Duschl, Heidi A. Schweingruber, & Andrew W. Shouse (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.10.17226/11625 Retrieved from

<https://www.nap.edu/catalog/11625/taking-science-to-school-learning-and-teaching-science-in-grades>

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York: NY: Routledge.

Sarama, J., & Clements, D. H. (2013). Lessons learned in the implementation of the TRIAD scale-up model: Teaching early mathematics with trajectories and technologies. In T. Halle, A. J. R. Metz & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 173-191). Baltimore, MD: Paul H. Brookes.

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, 1(2), 89-119. 10.1080/19345740801941332 Retrieved from  
<http://dx.doi.org/10.1080/19345740801941332>

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2016). Professional development in early mathematics: Effects of an intervention based on learning trajectories on teachers' practices. *Nordic Studies in Mathematics Education*, 21(4), 29-55.

Seago, N., Jacobs, J., & Driscoll, M. (2010). Transforming middle school geometry: Designing professional development materials that support the teaching and learning of similarity. *Middle Grades Research Journal*, 5(4), 199-211.

Seago, N., Jacobs, J., Driscoll, M., Nikula, J., Matassa, M., & Callahan, P. (2013). Developing teachers' knowledge of a transformations-based approach to geometric similarity. *Mathematics Teacher Educator*, 2(1), 74-85. 10.5951/mathtaceduc.2.1.0074

- Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 623-646) ACE/Praeger.
- Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26(2), 114-145.
- Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology*, 37(1), 57-86. 10.1146/annurev.soc.012809.102657  
Retrieved from <https://doi.org/10.1146/annurev.soc.012809.102657>
- Son, J. (2013). How preservice teachers interpret and respond to student errors: Ratio and proportion in similar rectangles. *Educational Studies in Mathematics*, 84(1), 49-70.  
10.1007/s10649-013-9475-5 Retrieved from <https://doi.org/10.1007/s10649-013-9475-5>
- Szilagyi, J., Clements, D. H., & Sarama, J. (2013). Young children's understandings of length measurement: Evaluating a learning trajectory. *Journal for Research in Mathematics Education*, 44(3), 581-620. 10.5951/jresematheduc.44.3.0581 Retrieved from <http://www.jstor.org/stable/10.5951/jresematheduc.44.3.0581>
- Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction toward a theory of teaching. *Educational Researcher*, 41(5), 147-156.
- Templin, J. (2015, June 30). Jonathan templin's website: Psychometrics and statistics as a way of life. Retrieved from [https://jonathantemplin.com/files/irt/irt15icpsr/irt15icpsr\\_lecture06.pdf](https://jonathantemplin.com/files/irt/irt15icpsr/irt15icpsr_lecture06.pdf)
- Usiskin, Z. P. (1972). The effects of teaching euclidean geometry via transformations on student achievement and attitudes in tenth-grade geometry. *Journal for Research in Mathematics Education*, 3(4), 249-259. 10.2307/748492 Retrieved from <https://doi.org/10.2307/748492>
- Vinner, S., & Hershkowitz, R. (1983). On concept formation in geometry. *Zentralblatt Fur Didaktik Der Mathematik*, 83(1), 20-25.

- Webel, C., Krupa, E., & McManus, J. (2015). Teachers' evaluations and use of web-based curriculum resources to support their teaching of the common core state standards for mathematics. *Middle Grades Research Journal*, 10(2), 49-64.
- Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning*, 13(4), 259-291. 10.1080/10986065.2011.608344 Retrieved from <http://dx.doi.org/10.1080/10986065.2011.608344>
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181-208. 10.1207/S15324818AME1302\_4 Retrieved from [https://doi.org/10.1207/S15324818AME1302\\_4](https://doi.org/10.1207/S15324818AME1302_4)
- Wilson, P. H., Sztajn, P., Edgington, C., & Confrey, J. (2014). Teachers' use of their mathematical knowledge for teaching in learning a mathematics learning trajectory. *Journal of Mathematics Teacher Education*, 17(2), 149-175. 10.1007/s10857-013-9256-1 Retrieved from <https://doi.org/10.1007/s10857-013-9256-1>
- Wilson, P. H., Mojica, G. F., & Confrey, J. (2013). Learning trajectories in teacher education: Supporting teachers' understandings of students' mathematical thinking. *The Journal of Mathematical Behavior*, 32(2), 103-121.
- Zucker, S., Sassman, C., & Case, B. (2004). *Cognitive labs*. (). San Antonio, TX: Pearson.

## **APPENDICES**

## Appendix A: IRB Approval Letter

5/17/2018

North Carolina State University Mail - Confrey - 7988 - IRB Protocol renewal/amendment approved

**NC STATE**

William McGowan <wgmcgowa@ncsu.edu>

---

### Confrey - 7988 - IRB Protocol renewal/amendment approved

---

**IRB Administrative Office** <pins\_notifications@ncsu.edu>  
Reply-To: debra\_paxton@ncsu.edu  
To: wgmcgowa@ncsu.edu

Thu, Sep 21, 2017 at 10:40 AM

Dear William McGowan:

Date: 09/21/2017

Project Title: Building a Next Generation Diagnostic Assessment and Reporting System within a Learning Trajectory-Based Mathematics Learning Map for Grades 6-8

IRB#: 7988

PI: Confrey, Jere

Approval period ends: 05/25/2018

The renewal/amendment request for the project listed above has been approved in accordance with policy under 45 CFR 46. If your application was to amend your study protocol, and your study received expedited or full board review, this letter does NOT change the expiration date for your study. If you applied to renew your expedited or full board protocol, your new expiration date is shown above.

1. This board complies with requirements found in Title 45 part 46 of The Code of Federal Regulations. For NCSU the Assurance Number is: FWA00003429.
2. You must use the approved documents which have the status "approved" in the document viewer in the eIRB for your study.
3. Any changes to the protocol and supporting documents must be submitted and approved by the IRB prior to implementation via amendment request.
4. If any unanticipated problems or adverse events occur, they must be reported to the IRB office within 5 business days by completing and submitting the unanticipated problem form on the IRB website: <http://research.ncsu.edu/sparcs/compliance/irb/irb-forms/>
5. Any unapproved departure from your approved IRB protocol results in non-compliance. Please find information regarding how to avoid non-compliance here: [http://research.ncsu.edu/sparcs-docs/irb/non-compliance\\_faq\\_sheet.pdf](http://research.ncsu.edu/sparcs-docs/irb/non-compliance_faq_sheet.pdf)

Please let us know if you have any questions..

Sincerely,

Deb Paxton  
919.515.4514  
IRB Administrator  
[dapaxton@ncsu.edu](mailto:dapaxton@ncsu.edu)  
NC State IRB Office

Jennie Ofstein  
919.515.8754  
IRB Coordinator  
[irb-coordinator@ncsu.edu](mailto:irb-coordinator@ncsu.edu)  
NC State IRB Office

[https://mail.google.com/mail/u/0/?ui=2&ik=186df13eb0&jsver=GAFHaMvshdw.en.&cbl=mailto\\_fe\\_180506.06\\_p7&view=pt&msg=15ea4e217f768acb&q=irb%20approval&qs=true](https://mail.google.com/mail/u/0/?ui=2&ik=186df13eb0&jsver=GAFHaMvshdw.en.&cbl=mailto_fe_180506.06_p7&view=pt&msg=15ea4e217f768acb&q=irb%20approval&qs=true)

## **Appendix B: Recruitment Letter and Parent Consent Forms**

Dear Parents:

This letter is to introduce you to a specific study that will be conducted by one of my graduate students for her doctoral dissertation and invite you to enroll your child in this study. Our research group is called SUDDS (Scaling Up Digital Design Studies) and your child's school (Highland Middle School) has been research partners with us since 2015. We are a team of learning scientists, mathematics educators, measurement specialists, designers, and software developers from NC State University working to help students enjoy math more, and be better prepared for high school mathematics.

I invite your child to participate in one of the critical phases of the project. Your child and others in her/his grade will be asked to take an assessment that is embedded within our digital learning system called *Math-Mapper 6-8*. Based on the results from this test, we intend to select a sample of 18 students to participate in think-aloud interviews in order to provide feedback to the educational researchers about the accuracy of the assessments. If your child is selected for the think-aloud interviews s/he will also be video recorded as s/he participates in the interview. Video recordings will be retained indefinitely on secure servers or secured hard drives by the PI and research team, and will be used in future research studies.

This project aims to improve students' learning experiences and their enjoyment of both mathematics and technology, so we take the students' feedback very seriously, and they get to see the changes we make as the result of *their* feedback. We consider the students will be our design partners in improving *Math-Mapper*.

We have included more information on the project in a consent form to be signed if you and your child choose to participate and be videotaped. Please return the consent form to \_\_\_\_\_ by \_\_\_\_\_ . If you have any questions at all, please contact us. We will be happy to answer any questions you may have.

We hope your child will be able to join us in this exciting new project.

Sincerely,

Jere Confrey, Ph. D  
Joseph D. Moore Distinguished Professor of Mathematics Education,  
NC State University

**North Carolina State University**  
**INFORMED CONSENT FORM for RESEARCH**

Title of Study: Applying a Validity Argument Framework to Learning Trajectories on Middle Grades Similarity Using Learning Science and Psychometric Lenses

Principal Investigator Dr. Jere Confrey

---

**What are some general things you should know about research studies?**

We invite your child to take part in a research study. Your child's participation in this study is voluntary; your child has the right to be a part of this study, to choose not to participate, or to stop participating at any time. The purpose of research studies is to gain a better understanding of a certain topic or issue. Your child is not guaranteed any personal benefits from being in a study. Some research studies also may pose risks to those that participate. In this consent form you will find specific details about the research in which your child is being asked to participate. If you do not understand something in this form it is your right to ask the researcher for clarification or more information. A copy of this consent form will be provided to you. If at any time you or your child have questions about your child's participation, do not hesitate to contact the researcher named above.

**What is the purpose of this study?**

Overall, our research project aims to understand better how children learn middle grades math topics so that we can develop valid assessments, within our digital learning system *Math-Mapper 6-8*, for students and teachers to use in the classroom.

This particular study aims to develop deeper understanding of how eighth-grade students think about geometrical similarity and if the digitally administered assessment items associated to geometrical similarity are accurately measuring student thinking. During the study, your child will be expected to use a combination of paper and pencil tools and electronic devices such as tablet computer (iPads) and laptop computers.

**What will happen if you take part in the study?**

If you and your child agree that your child will participate in this study, your child will be one of many in her/his grade who will first be asked to take an assessment that is embedded within our digital learning system called *Math-Mapper 6-8*. Your child will take this assessment during her/his regular math lessons. Based on the results from this test, we intend to select a sample of 18 students to participate in think-aloud interviews in order to provide feedback to the educational researchers about the accuracy of the assessments.

If your child is selected for the think-aloud interviews s/he will also be video recorded as s/he participates in the interview. Think-aloud interviews help us understand whether the assessment items are appropriate for the students and, and will give insight into whether the software is providing a learning structure that is useful to students and teachers. Such videotapes are important for doing the research: they help us reconstruct all the children's activities and responses during their work. Because in our research we often refer to these kinds of video recordings again and again, we plan to keep the recordings indefinitely for use in other research studies, and if you provide permission, excerpts of the recordings may be used for publication purposes. Video recordings will be retained indefinitely on secure servers or secured hard drives by the PI and research team, and will be used in future research studies.

Interviews will be conducted so as not to interfere with your child's required school activities; so, for instance, they may be scheduled after regular school hours, or at times such as intersession when students are not in school. In all cases, interviews will be scheduled through consultation with students and teachers or parents, as appropriate.

Each interview episode will typically require about an hour. We will communicate with you and your child ahead of time so that you and the research team can work out the scheduling. The interviews or small-group or class sessions will take place at your child's school. The interviews may take place over several sessions during the fall semester.

**Risks**

We do not foresee any risks of this research for your child. The particular mathematical problems your child may work on will be similar or identical to tasks that he or she might be asked to do in school. None of the students'

work on this project will have any consequence for their regular classes or grades. Assessment items they engage with will have no bearing on their school grades.

### **Benefits**

There are not necessarily any *direct* benefits for your child from participating in this research. However, there are a number of potential indirect benefits. Children we have worked with in other math learning and technology projects are excited to participate in developing new computer tools. We try at all times to make the tasks, including and especially any assessment items, stimulating and enjoyable for the students. In addition, because the children are encouraged to “think out loud,” and make their thinking visible with drawings and sketches, the activities may help students understand their own thinking better, reinforce their recognition that they do have good ideas, and that they are “better at math” than they might have thought. The students may broaden their notions of what math is, and become more confident in their own mathematical learning. Moreover, their experience in user testing may give them insight into an important phase of software technology development, and increase their interest in that field as an eventual career.

The project your child will participate in is a new innovative approach to improve mathematics learning in North Carolina. Your student will be part of research that, we hope, will directly benefit future students and teachers in learning and teaching mathematics, by supporting better teaching and testing that gives important feedback on student progress and helping teachers improve their teaching of mathematics for each student. Your child’s participation will help us gain insight into how to make mathematics instruction more effective in all classrooms.

### **Confidentiality**

The information in the study records will be kept confidential to the extent allowed by law. The research team will retain all digital data and non-digital work products the subjects generate (drawing, paper-and-pencil solutions/attempts, and other artifacts). All data (including video interviews) will be stored electronically on file storage leased by NC State University. Access is restricted by user name and password. With your consent (signature below) particular clips from the video recordings may be shared with people beyond our project team (for instance, professional meetings and teacher professional development classes or materials for teachers.). Project reports will include individual student responses to assessments items, strategies with and reactions to the software on the electronic devices. Children will not be identified by name in such presentations or materials, or in any written or oral reports of the research.

### **Compensation**

Your child will receive a \$15 gift card if s/he is selected for the think-aloud interviews.

### **What if you have questions about this study?**

If you have questions at any time about the study or the procedures, you may contact the researcher, Dr. Jere Confrey, at [jconfre@ncsu.edu](mailto:jconfre@ncsu.edu), or 919-306-8523.

### **What if you have questions about your rights as a research participant?**

If you feel you have not been treated according to the descriptions in this form, or your rights as a participant in research have been violated during the course of this project, you may contact Deb Paxton, Regulatory Compliance Administrator at [dapaxton@ncsu.edu](mailto:dapaxton@ncsu.edu) or by phone at 1-919-515-4514.

**Consent To Participate**

[Student]: I have read and understand the above information. I have received a copy of this form. I agree to participate in this study with the understanding that I may withdraw at any time.

[Parent or guardian]: I consent to my child's participation in this study, including videotaping of my child for research purposes. I affirm that I have discussed, with my child, my child's role in this study. I affirm that my child understands this, and has agreed to participate.

---

**Student Signature**

---

**Student Name (please print)**

---

**Grade**

---

**Parent Signature**

---

**Parent Name (please print)**

---

**Date**

---

**Parent Phone Number**

---

**Parent Email****Consent for Videotape Use in Presentations and Materials**

I additionally consent for short excerpts of video recordings of my child in this project, as judged useful by the research team, to be used for use in professional research presentations, teacher professional development materials. Recordings will be kept indefinitely and may be used in publications, as long as my child is not identified by name in such presentations.

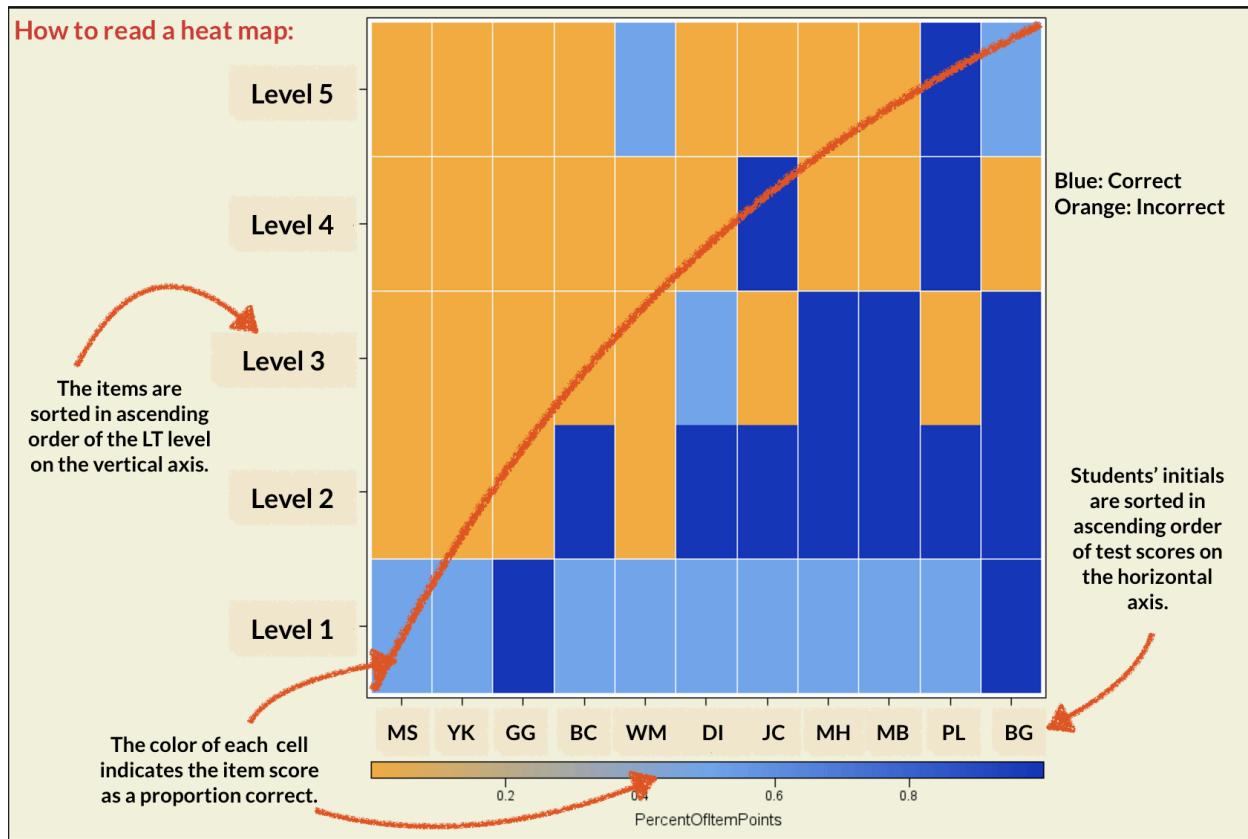
---

**Parent/Guardian Signature**

---

**Name of Student's School/Community Center**

## Appendix C: Example of a Teacher Report for One Construct



## Appendix D: Example of a Student Report

**My Results By Construct**

Construct	Percent Correct
Construct A: Similarity and Scaling	50%
Construct B: Transformations of Geometric Shapes	55%

**A: Similarity and Scaling**

Level	Description
L5	Defines similar triangles as having equal angles and sides in proportion and, given one pair of corr...
L4	Grows or shrinks a similar triangle by doubling, tripling, or halving side lengths and defines simil...
L3	Identifies similar triangles as having matching angles but possibly differing side lengths
L2	Identifies similar objects as having the "same shape" and justifies by informally scaling one onto the other

**Key:** █ Correct █ Revised Correctly █ Percent Correct: 100% █ 67-99% █ 34-66% █ 1-33% █ 0% █ Untested Level

**My Scores by Item and Construct**

Construct	1 Gr. 7 Level 3	2 Gr. 7 Level 4	3 Gr. 8 Level 9	4 Gr. 7 Level 6	5 Gr. 8 Level 2	6 Gr. 8 Level 3	7 Gr. 8 Level 3	8 Gr. 8 Level 4	9 Gr. 8 Level 5	My Construct Score
Construct A Similarity and Scaling	✓	✗	✓	✗						50%
	E: 1-2   7: 3-8   8: 9									
Construct B Transformations of Geometric Shapes					✓	✓	✓	✗	✗	55%
	B: 8: 1-6									

**Key:** ✓ Full Credit ▢ Partial Credit ✗ No Credit 🔖 Bookmarked  
✓ Full Credit After Retry ▢ Partial Credit After Retry ✗ No Credit After Retry 🚩 Misconception

Cluster Score: 53%

## Appendix E: Big Ideas in Math-Mapper 6-8

### GEOMETRY & MEASUREMENT

Measure, compose, and scale perimeter, area, and volume

Compose, characterize, and transform lines, angles, and polygons

Represent and explore Pythagorean Theorem and polygons using coordinate points

### ALGEBRA

Algebraically relate, express, modify, and evaluate unknown quantities

Represent and use relations and functions of two variables

### NUMBER

Position, compare, and operate on one dimensional quantities

Compare quantities to operate and compose with ratio, rate, and percent

### STATISTICS & PROBABILITY

Display data and use statistics to measure center and variation in distributions

Use probability to measure chance and model chance events to make informal inferences

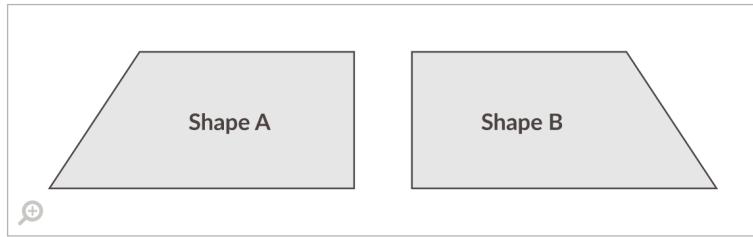
## Appendix F: Assessment Items aligned to Similarity and Scaling LT

Item ID 749

Level 1

Grade E

Five students were discussing whether the two shapes below are congruent. Their comments are in the table below.

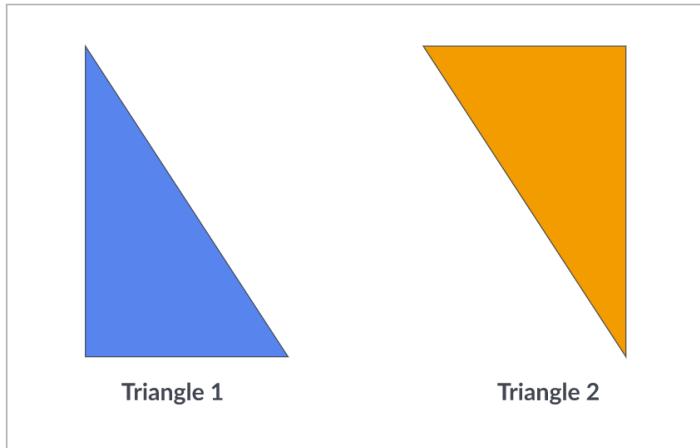


Which student is correct?

- Student 1
- Student 2 [1]
- Student 3
- Student 4
- Student 5
- None of these students are correct.

Student	Comment
Student 1	Shapes A and B are not congruent, because shape B is backwards.
Student 2	Shape A is congruent to shape B, because I can flip shape B and it fits exactly on top of shape A.
Student 3	Shapes A and B are congruent, because if I slide them together, they will fit into each other.
Student 4	Shapes A and B are not congruent, because they look different from each other.
Student 5	Shapes A and B are congruent, because they have the same area.

Consider the two triangles below.



Which statements about triangle 1 and triangle 2 are true? Check all that apply.

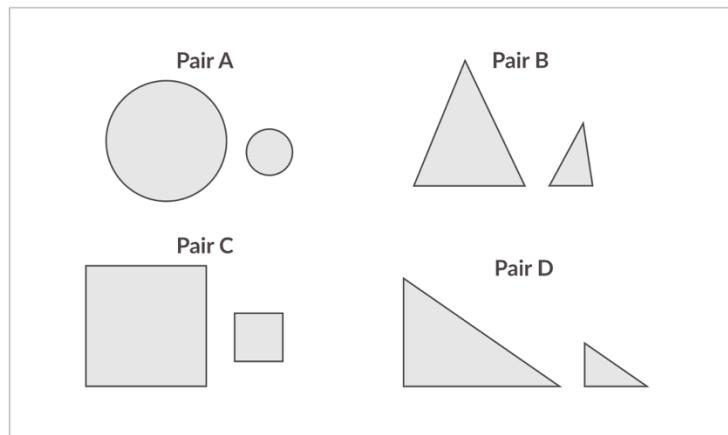
- Triangle 1 is congruent to triangle 2 if you can make triangle 2 fit perfectly on top of triangle 1. [
- Triangle 1 is not congruent to triangle 2 because triangle 2 is upside down.
- Triangle 1 is congruent to triangle 2 if they have the same side lengths and angle measurements
- Triangle 1 is not congruent to triangle 2 because they are different colors.
- Triangle 1 is congruent to triangle 2 if they have the same area.

Item ID 759

Level 2

Grade E

Below are four different pairs of shapes.



In which of these pairs could the larger shape be shrunk to become congruent to the smaller shape? Check all that apply.

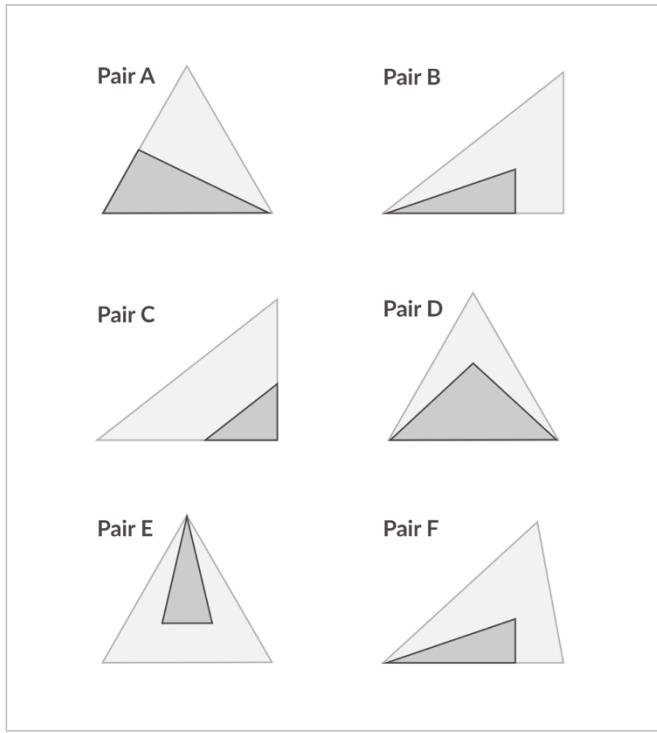
- Pair A [1]
- Pair B
- Pair C [1]
- Pair D [1]

Item ID 762

Level 2

Grade E

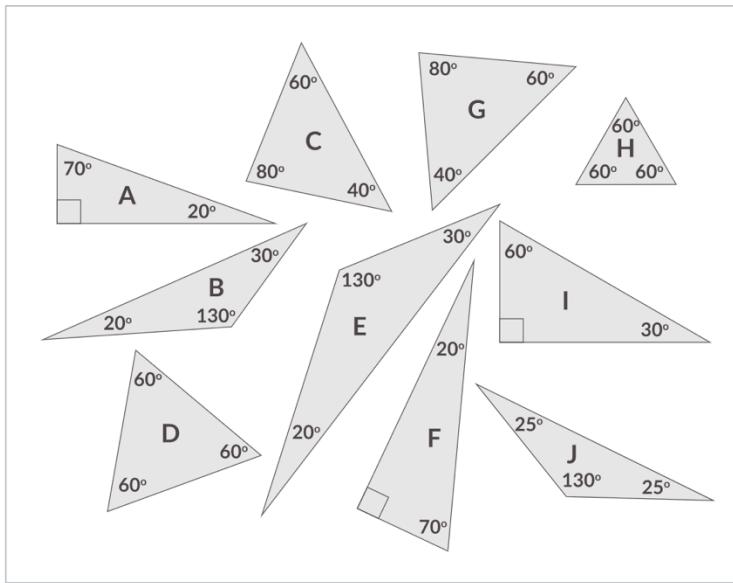
Carl wants to know which of the smaller triangular shapes could be enlarged on a photocopier to be congruent to the corresponding larger shape.



In which pairs, if any, are the triangles the same but different sizes? Check all that apply.

- Pair A
- Pair B
- Pair C [1]
- Pair D
- Pair E
- Pair F

Ten triangles are shown below.



For each of the following triangles,  
enter the letter of a similar triangle,  
other than itself. If no other triangle is  
similar to the given triangle, enter N.

Triangle A [1] [F]

Enter a letter

Triangle B [1] [E]

Enter a letter

Triangle C [1] [G]

Enter a letter

Triangle D [1] [H]

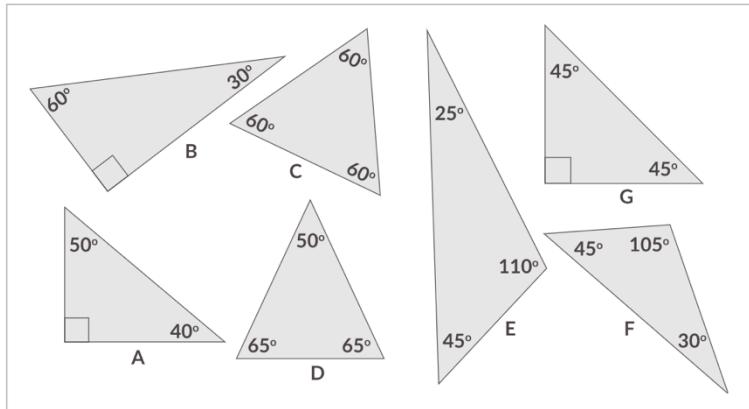
Enter a letter

Item ID 766

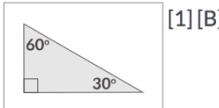
Level 3

Grade 7

Seven triangles are shown below.

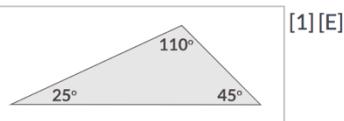


For each of the following triangles,  
enter the letter of a similar triangle.



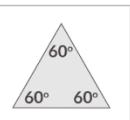
[1] [B]

Enter A, B, C, D, E, F, or G



[1] [E]

Enter A, B, C, D, E, F, or G



[1] [C]

Enter A, B, C, D, E, F, or G

Item ID 767

Level 3

Grade 7

Below are four pairs of triangles and statements about them.

Pair	Diagram	Statement
1		$\Delta ABC \sim \Delta FED$
2		$\Delta ABC \sim \Delta ADE$
3		$\Delta ABC \sim \Delta DEF$
4		$\Delta ABC \sim \Delta DEF$

Indicate whether the statement in the table about each pair of triangles is true (T) or false (F).

Pair 1 [1] [t]

Enter T or F

Pair 2 [1] [t]

Enter T or F

Pair 3 [1] [f]

Enter T or F

Pair 4 [1] [t]

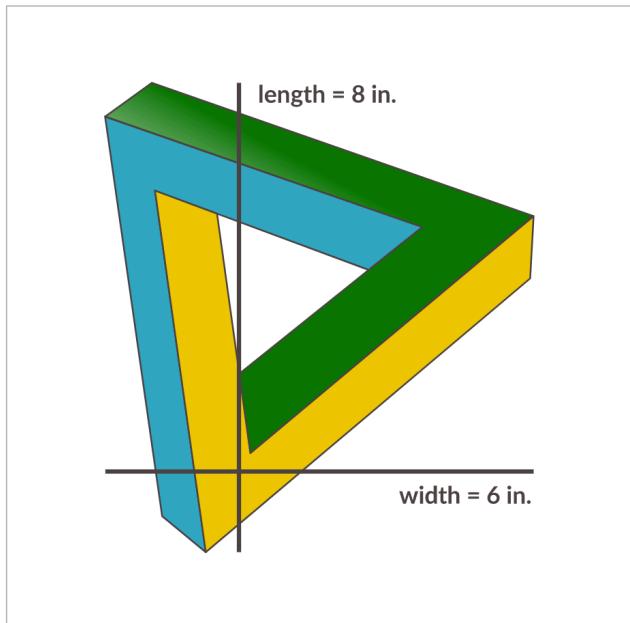
Enter T or F

Item ID 763

Level 4

Grade 7

Diedre wants to put the logo below on the back of her cell phone. She measures the logo and finds that it is 8 inches long by 6 inches wide. Her cell phone is 6 inches long by 3 inches wide.



Diedre knows that she needs to shrink the logo for it to fit on her phone. Which size logo will fit on her phone and not be distorted?

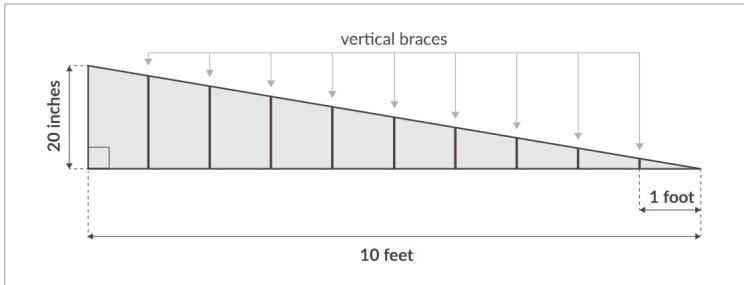
- 2 in. long by 3 in. wide
- 5 in. long by 3 in. wide
- 4 in. long by 3 in. wide [1]
- 6 in. long by 4 in. wide
- 6 in. long by 4.5 in. wide

Item ID 764

Level 4

Grade 7

Stuart's grandmother uses a wheelchair, and needs a handicap ramp to enter Stuart's house. Stuart's porch is 40 inches tall. Stuart wants to build his grandmother a handicap ramp that is the same steepness as the one at the hospital. The hospital's ramp has a vertical brace every foot. The hospital ramp's dimensions are shown below.



How far out into his yard would the ramp go (in feet) if it starts at the top of the porch and matches the steepness of the hospital ramp?

[1]  $20.0 \leq x \leq 20.0$

Enter a number

When a person in a wheelchair goes up the hospital ramp, each time she passes a brace, how many inches higher off the ground is she?

[1]  $2.0 \leq x \leq 2.0$

Enter a number

When a person in a wheelchair goes up Stuart's ramp, each time she passes a brace, how many inches higher off the ground is she?

[1]  $2.0 \leq x \leq 2.0$

Enter a number

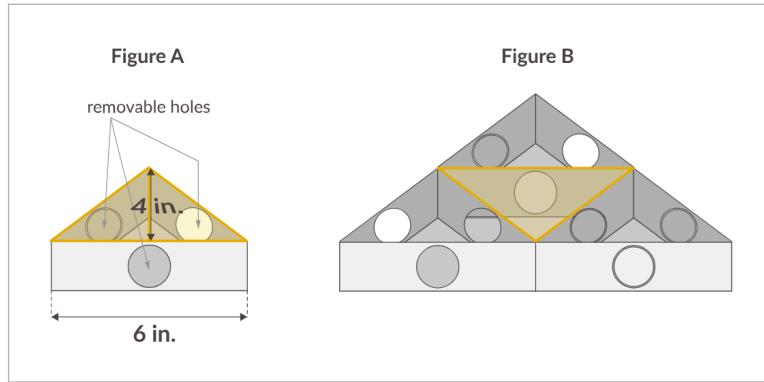
Item ID 768

Level 4

Grade 7

Tina invented a new way to build hamster cages. She designed a triangular plastic shape that could snap together with identical shapes to build larger hamster cages.

The plastic shape has a triangular top with a base of 6 inches and an altitude of 4 inches (see figure A below). Each plastic shape has three rectangular walls with removable circles to let the hamsters move between the plastic shapes. Using four plastic shapes, she built the large triangular hamster cage shown in figure B below.



nents/6\_19\_49\_04\_02\_aY1DB72.svg

How wide is the base of the large triangular top of figure B?

[1]  $12.0 \leq x \leq 12.0$

Enter a number

What is the altitude of the large triangular top of figure B?

[1]  $8.0 \leq x \leq 8.0$

Enter a number

Tina wants to make a larger triangular hamster cage with a 42-inch wide base.

How many triangular plastic shapes should she place along the base?

[1]  $7.0 \leq x \leq 7.0$

Enter a number

What would be the altitude of the top of this larger triangular cage?

[1]  $28.0 \leq x \leq 28.0$

Enter a number

How many triangular plastic shapes would this larger triangular cage have in total?

[1]  $49.0 \leq x \leq 49.0$

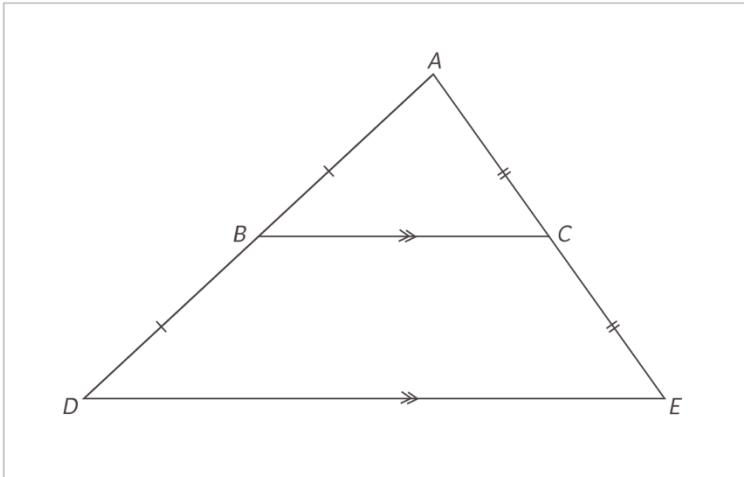
Enter a number

Item ID 760

Level 5

Grade 7

$B$  and  $C$  are the midpoints of  $DA$  and  $EA$ , respectively.  $BC$  is parallel to  $DE$ .  
Triangle  $ABC$  is similar to triangle  $ADE$ .



If the length of  $DE$  is 18, then what is the length of  $BC$ ?

[1]  $[9.0 \leq x \leq 9.0]$

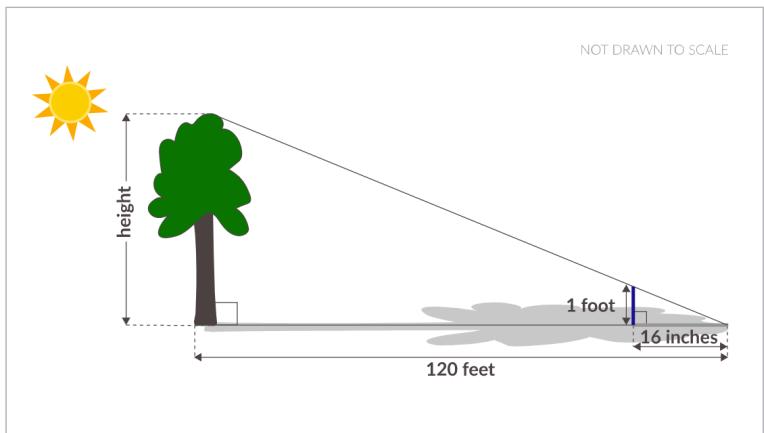
Enter a number

Item ID 769

Level 5

Grade 7

Clara's foot-long ruler makes a 16-inch shadow on a flat surface. At the same time, a tree lined up with Clara's ruler makes a shadow that is 120 feet long, as shown in the diagram below. Clara can use this information to determine the height of the tree.



How tall is the tree in feet? [1]  $[90.0 \leq x \leq 90.0]$

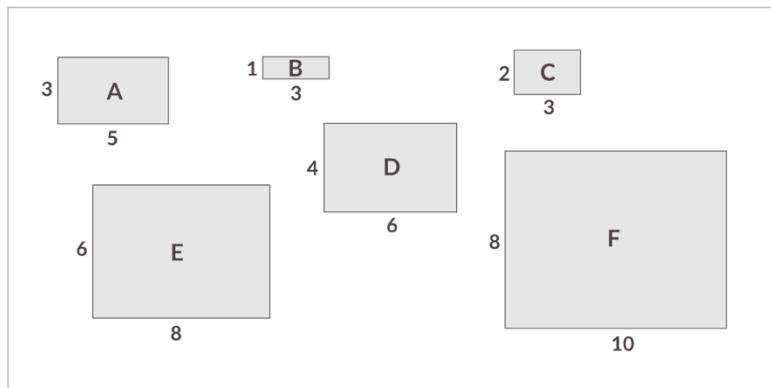
Enter a number

Item ID 750

Level 6

Grade 7

Below are six rectangles that are different sizes.



Which rectangles are similar to rectangle D? Check all that apply.

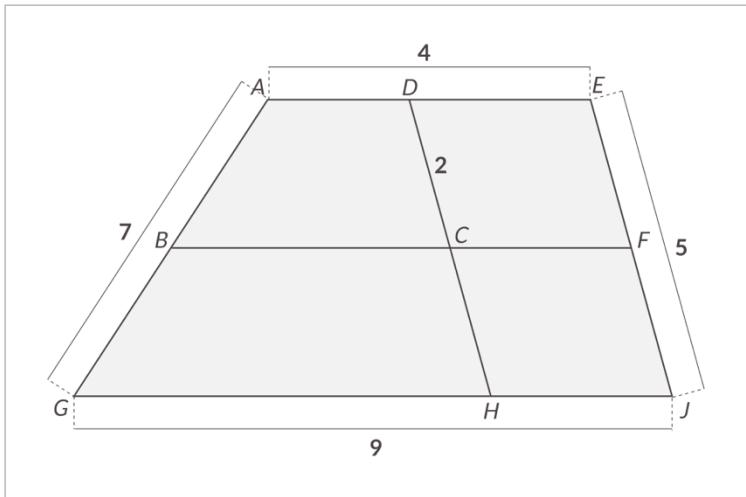
- A
- B
- C [1]
- E
- F

Item ID 771

Level 6

Grade 7

In the figure below, trapezoid  $ABCD$  is similar to trapezoid  $AGJE$ .



Find the length of each side of trapezoid  $ABCD$ . If necessary, round your answers to one decimal place.

The length of side  $AD$  is \_\_\_\_\_. [1]  $[1.6 \leq x \leq 1.6]$

Enter a number

The length of side  $BC$  is \_\_\_\_\_. [1]  $[3.6 \leq x \leq 3.6]$

Enter a number

The length of side  $AB$  is \_\_\_\_\_. [1]  $[2.8 \leq x \leq 2.8]$

Enter a number

Item ID 773

Level 6

Grade 7

Trevor has a photograph of his grandmother that is  $4'' \times 6''$ . He wants to enlarge it to fit in a larger frame with no white space around the photograph and no distortion.

Which frames could he use? Check all that apply.

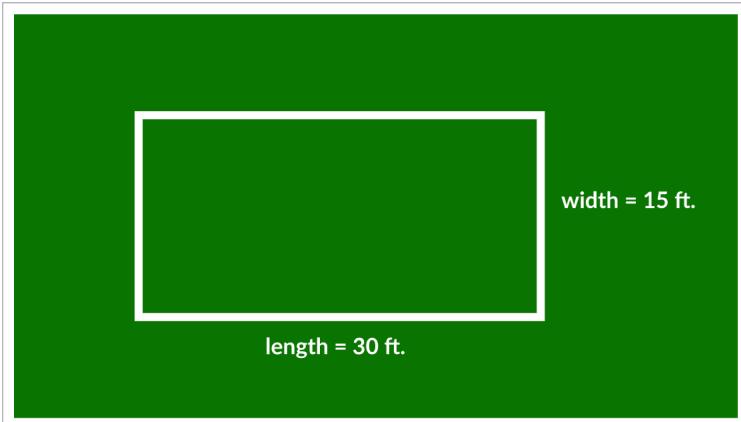
- $5'' \times 7''$
- $6\frac{2}{3}'' \times 10''$  [1]
- $8'' \times 10''$
- $8'' \times 12''$  [1]
- $10'' \times 15''$  [1]

Item ID 800

Level 6

Grade 7

Arthur painted boundaries for a rectangular soccer field in his backyard for his kids. The dimensions are shown below.



Arthur chose these measurements when his children were small, but now they are bigger. They asked him to make a bigger soccer field that has the same ratio of length to width as the one to the left. Which of the following measurements could he use to make this bigger field? Check all that apply.

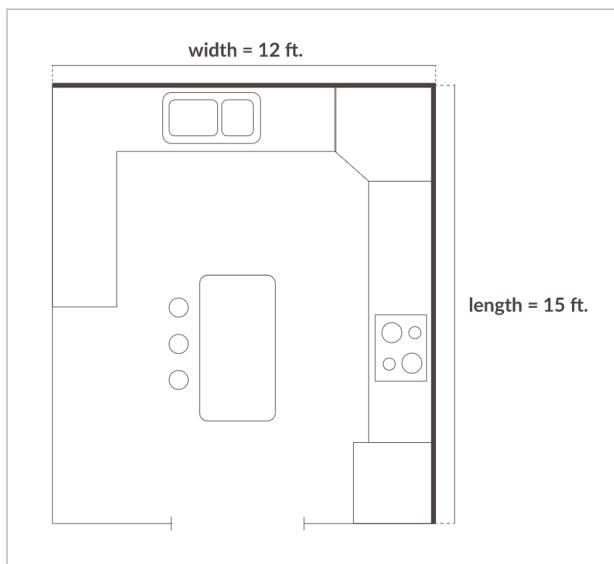
- Length = 30 ft. and width = 30 ft.
- Length = 60 ft. and width = 15 ft.
- Length = 60 ft. and width = 30 ft. [1]
- Length = 50 ft. and width = 25 ft. [1]

Item ID 751

Level 7

Grade 7

Chantelle's mom is remodeling the kitchen in her home. The actual dimensions of the kitchen are shown below. The designer she hired brought a scale drawing of the new kitchen for Chantelle's approval before construction.



The drawing is a  $\frac{1}{12}$ -scale model of the real kitchen. Complete the following sentences.

The actual length of the kitchen is 15 ft. The length of the kitchen in the scale drawing is \_\_\_\_ ft.

[1]  $[1.25 \leq x \leq 1.25]$

Enter a number

The actual width of the kitchen is 12 ft. The width of the kitchen in the scale drawing is \_\_\_\_ ft.

[1]  $[1.0 \leq x \leq 1.0]$

Enter a number

The actual area of the kitchen is \_\_\_\_ times as big as the area of the kitchen in the scale drawing.

[1]  $[144.0 \leq x \leq 144.0]$

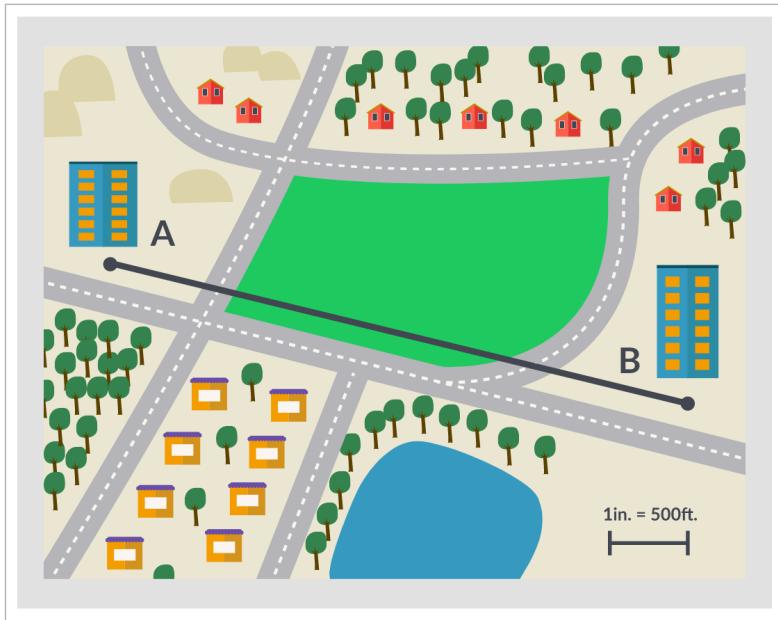
Enter a number

Item ID 770

Level 7

Grade 7

A map is drawn with a scale as shown, where 1 inch represents 500 feet.  
Buildings A and B are 7.2 inches apart on the map.



Find, in feet, the actual distance between the two buildings.

[1]  $3600.0 \leq x \leq 3600.0$

Enter a number

Item ID 236

Level 8

Grade 7

A hotel owner is replacing the tile in his ballroom. There are three sizes of tile to choose from, listed below.

Tile dimensions	Price
6 in. × 6 in.	?
12 in. × 12 in.	\$1 per tile
18 in. × 18 in.	?

How much would one 18 in. × 18 in. tile need to cost, in dollars, for the cost per square inch to be the same as the 12 in. × 12 in. tiles?

[1]  $2.25 \leq x \leq 2.25$

Enter a decimal number

How much would one 6 in. × 6 in. tile need to cost, in dollars, for the cost per square inch to be the same as the 12 in. × 12 in. tiles?

[1]  $0.25 \leq x \leq 0.25$

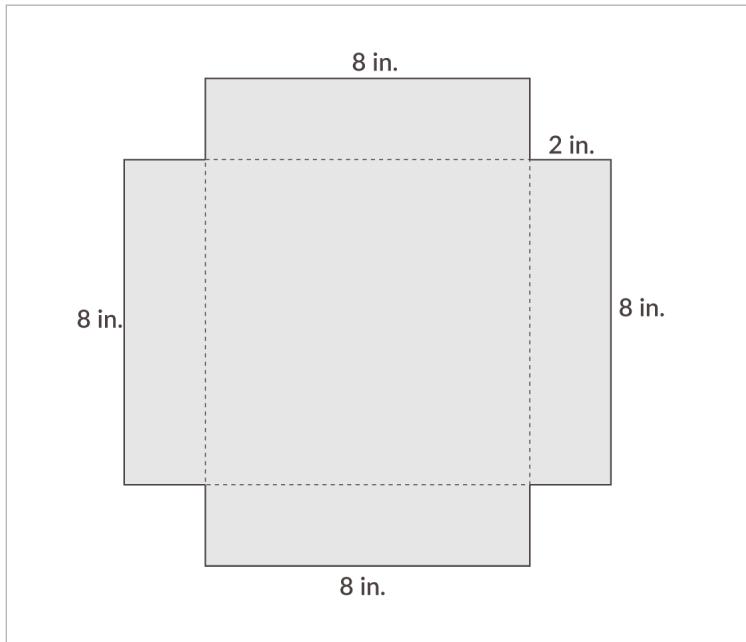
Enter a decimal number

Item ID 241

Level 8

Grade 7

Jen needs a box to hold a cake she is baking. She cuts a piece of cardboard into the shape below and folds the sides to make an open box.



The box is not large enough to hold the cake, so Jen decides to use a cardboard piece in the same shape, but she doubles each of the current dimensions. By doubling the dimensions of the cardboard box, how has Jen changed the volume of the box?

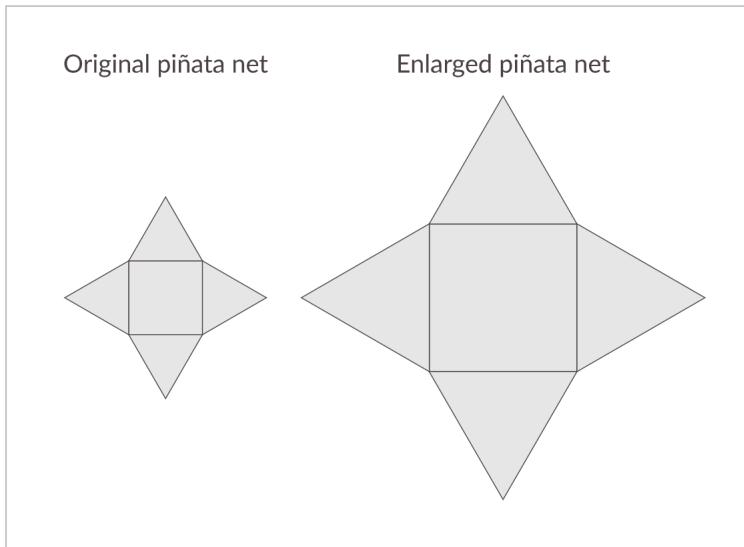
- The volume of the new box is 2 times the volume of the old box.
- The volume of the new box is 4 times the volume of the old box.
- The volume of the new box is 6 times the volume of the old box.
- The volume of the new box is 8 times the volume of the old box.

Item ID 433

Level 8

Grade 7

Chardae bought a rectangular pyramid to use as a piñata at a party. She wants to make a bigger piñata that looks exactly the same as the original, so she doubled the length of all the edges of the pyramid. Then she cut a new net out of cardboard with these new measurements. Both nets are shown below.



How much more cardboard will the enlarged piñata use than the original piñata?

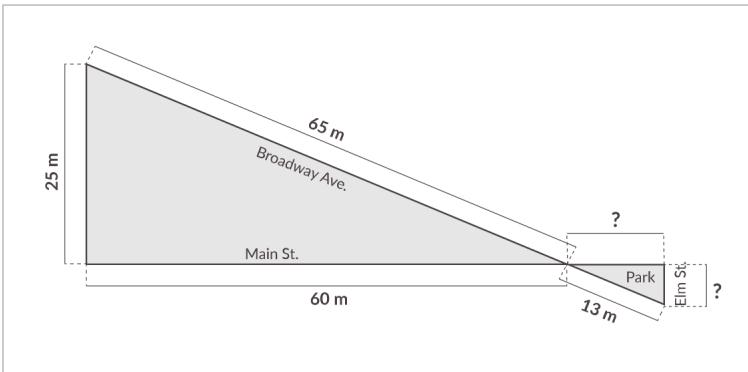
- 2 times as much cardboard
- 4 times as much cardboard [1]
- 8 times as much cardboard
- There is no way to know without the measurements of the piñatas.

Item ID 761

Level 9

Grade 8

An architect is designing a park for a small triangular piece of land formed by the intersection of Broadway Avenue and Main Street. She needs the dimensions of the piece of land. She knows the dimensions of the larger triangle, as shown, and she knows the length of the park along Broadway Avenue is 13 meters.



What is the length of the park along Main Street?

[1]  $12.0 \leq x \leq 12.0$

Enter a number

What is the length of the park along Elm Street?

[1]  $5.0 \leq x \leq 5.0$

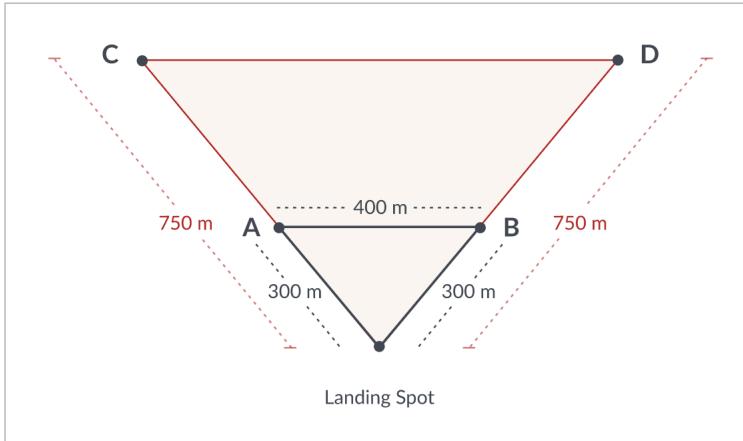
Enter a number

Item ID 1182

Level 9

Grade 8

Two rovers are exploring a planet in a nearby star system. They start at a landing spot and drive away in straight lines at the same speed. When the rovers are both 300 meters away from the landing spot at points A and B, they are 400 meters apart from each other. They continue driving in straight lines, as shown in the figure below.



Calc

When the two rovers are at points C and D (750 meters away from the landing spot), how far apart (in meters) are they from each other?

[1]  $1000.0 \leq x \leq 1000.0$

Enter a number

When the two rovers are 12,000 meters apart from each other, how far away are they (in meters) from the landing spot?

[1]  $9000.0 \leq x \leq 9000.0$

Enter a number

## **Appendix G: Interview Protocol**

Student Name: \_\_\_\_\_

Time: \_\_\_\_\_

Date: \_\_\_\_\_

### **Researcher/Interviewer introduces the student to the study:**

*Hi! Thank you for participating in the study. Just for your information, I am collecting data for my PhD study. The data collected today will not be shared with anyone outside my research team. I will not be sharing the data with your teachers. I also want you to know that I am not assessing you today. You are not getting a score for this, I am more interested in the thoughts that are going through your mind as you are attempting the question.*

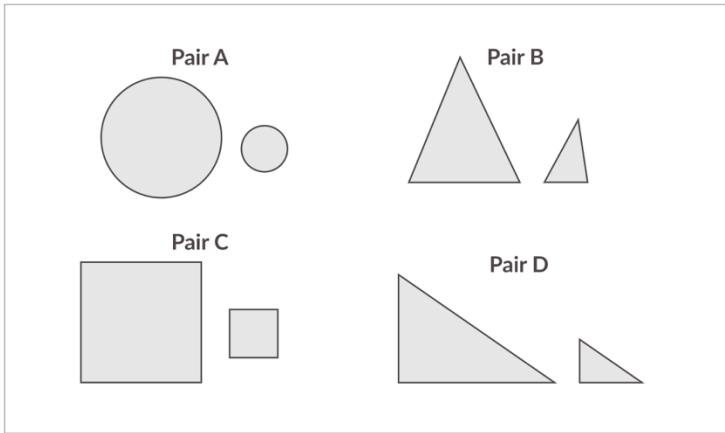
*I am going to describe today's interview session to you first. Please feel free to ask questions if you are not sure about the process. The interview is made up of two parts [Concurrent and Retrospective Verbalizations]. I will explain the parts to you. First, you will attempt a question using the SUDDS system, just as you would normally do when you're in school, but the only difference is that I would like you to say out loud whatever you are saying silently to yourself as you are answering a question. I will **not** interrupt you as you complete the question but I may keep reminding you to "keep talking" if you stop thinking-out-loud. Does that first part make sense? Do you have any questions? [Respond to student questions, if any]*

*In the second part, I may ask you some clarifying questions about why you responded in the way you did, or why you were thinking about something on the question in a certain way. I may ask these questions whether or not you answered the questions correctly. This second part is for me to make sure that I have understood your thought processes properly. Does that second part make sense? Do you have any questions? [Respond to student questions, if any]*

*We will carry out this process for a practice question, that way if you want to ask any more clarifying questions about the process you can do so before we start with the actual questions. There is a total of 7 questions that you will respond to in the same way described above, this includes the practice question. If at any point of the interview you feel uncomfortable and you want to take a break, or stop, just let me know. We will start with the practice question when you are ready. [Let the student get comfortable and indicate to you when they are settled in and ready to start].*

## **Practice Question**

Below are four different pairs of shapes.



In which of these pairs could the larger shape be shrunk to become congruent to the smaller shape? Check all that apply.

- Pair A [1]
- Pair B
- Pair C [1]
- Pair D [1]

**Make Sure Camera and Mic are on and recording.**

### **Instructions for Concurrent Verbalization:**

*Answer this question. Remember say out loud whatever you are saying silently to yourself as you attempt this question.*

### **Notes and Clarifying Questions for Retrospective Verbalization:**

**Check-in with the students if they have any questions about the interview process.**

*Okay, that was great. Do you feel comfortable? Are there any questions about the process? Let me know when you would like us to start the actual questions.*

**Announce the following information for the video recording:**

*Today is [date]. This is a think-aloud interview with [student name]. Let's begin with Question 1, for my reference this is Item ID [ID number].*

**Question:** 1

**Item ID:** \_\_\_\_\_

**Instructions for Concurrent Verbalization:**

*Answer this question. Remember say out loud whatever you are saying silently to yourself as you attempt this question.*

**Notes and Clarifying Questions for Retrospective Verbalization:**

## Appendix H: Think-aloud Distribution of Items across Student

<b>Theta-score on Similarity and Scaling Pre-test</b>	<b>Student Think-Aloud Numbers</b>
<b><math>-1.5 \leq \theta &lt; -0.35</math></b>	TA2, TA3, TA4, TA5, TA6
<b><math>-0.35 \leq \theta &lt; 0.74</math></b>	TA7, TA8a, TA9, TA10, TA11, TA12
<b><math>0.74 &lt; \theta &lt; 1.82</math></b>	TA8b, TA13, TA14, TA15, TA16, TA17, TA18

<b>ITEM</b>	<b>LEVEL</b>	<b>STUDENT THINK- ALOUD NUMBERS</b>
<b>749</b>	1	TA5, TA8a, TA8b, TA12, TA15
<b>1181</b>	1	TA4, TA7, TA11, TA14, TA18
<b>759</b>	2	TA5, TA8a, TA8b, TA12, TA15
<b>762</b>	2	TA2, TA4, TA9, TA11, TA16
<b>765</b>	3	TA3, TA6, TA10, TA13, TA17
<b>766</b>	3	TA2, TA7, TA9, TA14, TA16, TA18
<b>767</b>	3	TA3, TA6, TA10, TA13, TA17
<b>763</b>	4	TA4, TA9, TA16
<b>764</b>	4	TA2, TA5, TA10, TA14,

		TA16, TA18
<b>768</b>	4	TA3, TA6, TA11, TA17, TA18
<b>760</b>	5	TA7, TA12, TA14
<b>769</b>	5	TA2, TA8a, TA8b, TA11, TA13
<b>750</b>	6	TA3, TA7, TA12, TA15
<b>771</b>	6	TA4, TA8a, 8b, TA13, TA16
<b>773</b>	6	TA5, TA9, TA14, TA17
<b>800</b>	6	TA6, TA10, TA15, TA18
<b>751</b>	7	TA5, TA8a, TA8b, TA11, TA13
<b>770</b>	7	TA2, TA6, TA9, TA12, TA14, TA18
<b>236</b>	8	TA3, TA8a, TA8b, TA10, TA16
<b>241</b>	8	TA2, TA4, TA9, TA12, TA15, TA17
<b>433</b>	8	TA5, TA7, TA11, TA12, TA15, TA18

<b>761</b>	9	TA3, TA6, TA10, TA14, TA17
<b>1182</b>	9	TA4, TA7, TA13, TA15, TA16

**Appendix I: Independent Assessment of Geometric Similarity**

**7 G A 1 Similarity and Scaling Independent Assessment**

**Instructions:**

- 1. Circle the correct answer choice for each question.**
- 2. You may do your working out in the blank space within each question.**

Student Name:

---

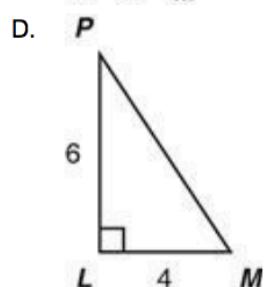
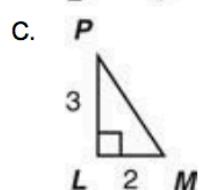
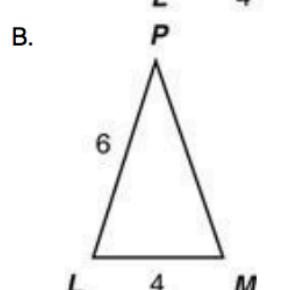
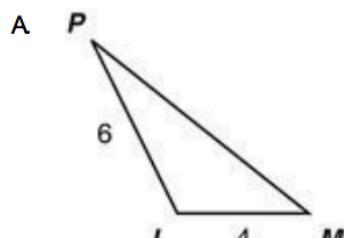
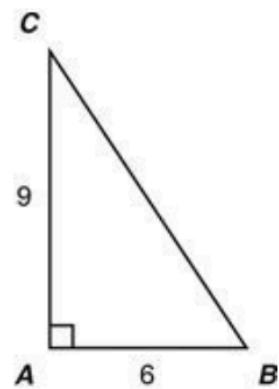
Class:

---

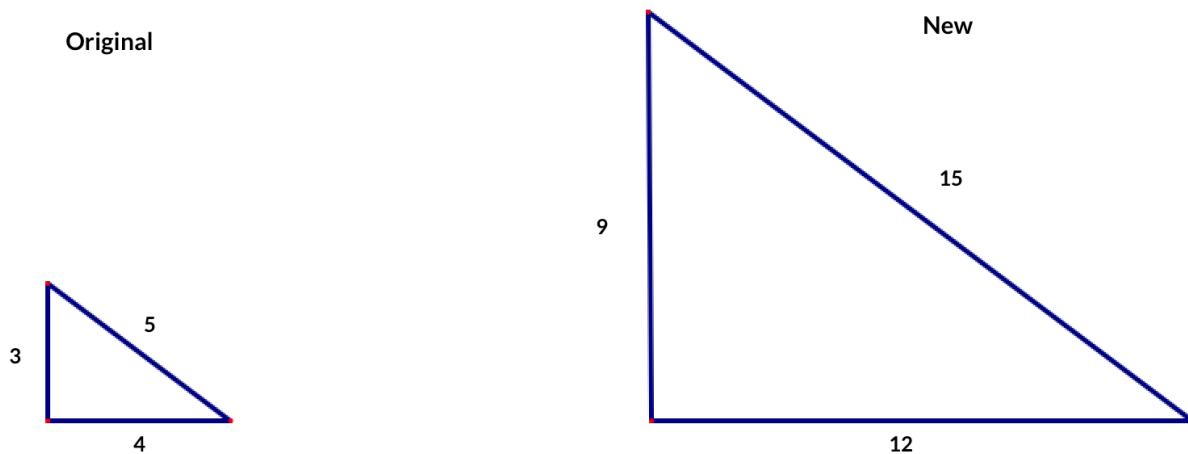
Date:

---

- 1. Which is a dilation of  $\Delta ABC$  with the scale factor of  $\frac{2}{3}$ ?**



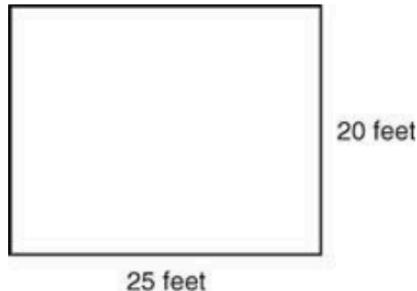
2. Observe the two triangles.



**How were the original dimensions changes to obtain the new dimensions?**

- A. 6 was added to all of the original dimensions.
  - B. 8 was added to all of the original dimensions.
  - C. The original dimensions were all multiplied by 3.
  - D. The original dimensions were all multiplied by 5.
- 

3. Mr. Stewart is making a scale drawing of his classroom. The actual dimensions of the room are shown below.

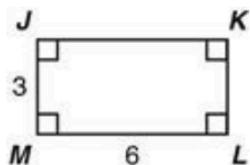


**Which dimensions could he use in the scale drawing?**

- A. 11 inches  $\times$  6 inches

- B. 10 inches  $\times$  8 inches
- C. 12 inches  $\times$  9 inches
- D. 15 inches  $\times$  10 inches

4. Rectangles  $JKLM$  is shown.



Which figure is a dilation of Rectangle  $JKLM$  with a scale factor of  $\frac{2}{3}$ ?

- A. 

3  
9
- B. 

2  
6
- C. 

3  
4
- D. 

2  
4

- 5. An architect drew a 1:15 scale model of a building. What statement most accurately describes what 1:15 represents in the drawing?**
- A. The drawing of the building is 15 inches long.
  - B. The actual building has a wall that measures 15 feet long.
  - C. 15 inches on the drawing represents 115 inches on the actual building.
  - D. One inch on the drawing represents 15 inches on the actual building.
- 
- 6. Chris has a miniature model of the boat that he is building. The length of the model is 4.5 inches and height is 2 inches. If the length of the actual boat is 18 feet, what is the height?**
- A. 8 feet
  - B. 9 feet
  - C. 18 feet
  - D. 36 feet
- 7. On a certain state map,  $\frac{1}{2}$  inch equals an actual distance of 10 miles. The distance on the map between City *R* and City *M* is  $5\frac{1}{2}$  inches. What is the actual distance between the two cities?**
- A. 55 miles
  - B. 100 miles
  - C. 105 miles
  - D. 110 miles
- 
- 8. Mrs. Townsend wants to draw a map of California for her bulletin board. The distance from the bottom of the state to the top is about 655 miles. As shown below.**



If Mrs. Townsend makes a scale drawing of the map on a sheet of paper that is 11 inches in height, which ratio will be MOST appropriate for her scale?

- A. 1 mile : 60 inches
  - B. 1 mile : 6000 inches
  - C. 1 inch : 60 miles
  - D. 1 inch : 600 miles
- 

9. Lawrence wants to enlarge a picture of his dog and put it in a frame. The dimensions of the original picture are 5 inches by 7 inches. The perimeter of the frame is 48 inches. The dimensions of the frame are proportional to the dimensions of the original picture. What should be the dimensions of the enlarged picture?

- A. 6 in. by 8 in.

- 
- B. 10 in. by 14 in.
  - C. 12 in. by 12 in.
  - D. 20 in. by 28 in.

---

**10. A drawing of car has a scale factor of 1 inch (in.) = 2 feet (ft.). A photograph of the same car is  $\frac{1}{5}$  the size of the drawing. What is the length of the car window in the photograph is the length of the actual car window is 2.5 feet?**

- A. 0.25 in.
  - B. 0.4 in.
  - C. 10 in.
  - D. 12.5 in.
- 

**11. Two students created models of the same building. One student used a scale factor of 3 centimeters (cm) = 15 meters (m). The second student used a scale factor of 4.5 cm = 15 m. the height of the original building is 20 meters. What is the difference between the heights if the two students' models?**

- A. 1.5 cm
  - B. 2 cm
  - C. 6.5 cm
  - D. 33 cm
-

**12. Carla makes a scale drawing of her garden that is 30 feet long and 24 feet wide. If the garden in her drawing is 20 inches long, what is its width, in inches?**

- A. 14
  - B. 16
  - C. 25
  - D. 26
- 

**13. On a scale drawing for a house, the dining room is 3 inches by  $3\frac{1}{2}$  inches.**

- The scale for the drawing is  $\frac{1}{2}$  inch = 2 feet.
- Flooring costs \$ 2.75 per square foot.

**How much will it cost to put new flooring in the dining room?**

- A. \$115.50
  - B. \$143.00
  - C. \$303.18
  - D. \$462.00
- 

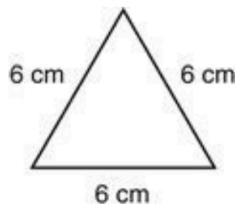
**14. Timothy has a drawing of the family room of a house that he is building.**

- The drawing shows the family room to be a rectangle 4 cm wide by 6 cm long.
- The length of the actual family room is 9 m.

What is the actual area of the family room?

- A.  $6 \text{ m}^2$
  - B.  $13.5 \text{ m}^2$
  - C.  $54 \text{ m}^2$
  - D.  $121.5 \text{ m}^2$
- 

**15. The triangle has a perimeter of 18 centimeters and an area of 93 square centimeters.**



**How does the area of a similar triangle with 3 times the perimeter compare to the area of the triangle shown?**

- A. It is 3 times as large.
- B. It is 6 times as large.
- C. It is 9 times as large.
- D. It is 18 times as large.

---

END OF TEST