

ROLE OF LINEAR REPRESENTATION OF LARGE MAGNITUDES ON  
UNDERSTANDING AND ESTIMATION

---

A dissertation  
Submitted to the Temple University Graduate Board

---

In Partial Fulfillment of the Requirements for the Degree of  
DOCTOR OF PHILOSOPHY

---

By  
Ilyse Resnick  
May, 2013

Examining Committee Members:

Thomas F. Shipley, Department of Psychology

Nora Newcombe, Department of Psychology

Peter Marshall, Department of Psychology

Elizabeth Gunderson, Department of Psychology

Christine Massey, Department of Psychology University of Pennsylvania

Alexandra Davatzes, Department of Earth Sciences

## ABSTRACT

Having a linear representation of magnitude across scales is essential in understanding many scientific concepts (Tretter, et al., 2006a) and is predictive of a range of mathematical achievement tests (Siegler & Booth, 2004). Despite the importance of understanding magnitude and scale, people have substantial difficulty comparing magnitudes outside of human perception (e.g., Jones, et al., 2008). The present work aims to examine the way people learn to represent and reason about large magnitudes through the development of two science of learning activities based on hierarchical alignment activity and corrective feedback.

The hierarchical alignment activity utilizes several analogical reasoning principles: hierarchical alignment, progressive alignment, structural alignment, and multiple opportunities to make analogies. Study 1 examines the effectiveness of hierarchical alignment by contrasting it with a conventional activity that uses all the analogical reasoning principles described above except for hierarchical alignment. Study 2 examines a corrective feedback activity, based on the same analogical reasoning principles used in study 1, except, using corrective feedback instead of progressive alignment and hierarchical alignment. Thus, study 2 examines the necessity of hierarchical and progressive alignment.

That both activities were successful in developing linear representations of geologic time (and for study 1, astronomical distances), suggests that multiple opportunities to make analogies through structural alignment are key components in developing analogies for learning magnitude. There appears to be an additive benefit of including hierarchical alignment (i.e., practice aligning magnitude relations across scales)

in analogies for learning about magnitudes. Corrective feedback may also be a useful strategy in learning about scale information. Pedagogical implications are discussed.

Both activities were based on the hypothesis that magnitudes at scales outside human perception are represented and reasoned about in the same way as magnitudes at human scales. The Category Adjustment Model (Huttenlocher, et al., 1988) suggests magnitude at human scales is stored as a hierarchical combination of metric and categorical information. People may use category boundaries to help make estimations in lieu of precise metric information. Variation in estimation, therefore, occurs because of imprecision of category boundaries (Shipley & Zacks, 2008; Zacks & Tversky, 2001). The current studies provided salient category boundaries to develop a more linear representation of magnitude. Thus, the effectiveness of the hierarchical alignment activity and the corrective feedback activity supports the hypothesis that people use hierarchically organized categorical information when making estimations across scales and across dimensions; and that providing people with more salient category boundary information improves estimation. Similarities and differences among temporal, spatial, and abstract line estimations are identified. Theoretical implications, including the potential application of the Category Adjustment Model to mental number lines, are discussed.

## ACKNOWLEDGEMENTS

I am extremely thankful for the wonderfully supportive environment I experienced in graduate school. This dissertation would not have been possible without the help, support, guidance, and efforts of many people.

In particular, I would like to thank my advisor, *Thomas F. Shipley*, for all his time and effort in helping to develop me as a researcher. I especially appreciate our lengthy conversations in the lab (both on task and tangential), your patience with my writing, and you affording me the unexpected opportunity to immerse myself in geology.

*Nora Newcombe*, thank you for your time and help throughout my graduate studies. In particular, thank you for your mindfulness of my career goals as I learn to navigate working at the intersection of cognitive science and geoscience.

*Alexandra Davatzes*, thank you for your temporal and geologic insights, and your work in developing the corrective feedback activity. It has been a lot of fun working with you.

*Christine Massey*, thank you for your help in developing my research, as well as the numerous article suggestions as I began thinking about new fields.

I would like to thank the members of my dissertation committee: *Alexandra Davatzes*, *Elizabeth Gunderson*, *Christine Massey*, *Peter Marshall*, *Nora Newcombe*, and *Thomas F. Shipley* for all their time and help on this project.

I would also like express gratitude to the geologists who have done a great deal to develop my geologic knowledge, such as *Alexandra Davatzes, Nicholas Davatzes, Tom Hickson, Kim Kastens, Cathy Manduca, Carol Ormond, and Basil Tikoff*. These geologists have invited me on their field trips, into their classrooms, to talks and conferences, as well as answered countless questions and facilitated data collection with expert geologists. Your generosity has been overwhelming. I would also like to thank the *geology instructors* who allowed me access to their classrooms to collect data.

I would like to thank my *fellow graduate students and lab members*. The opportunity to give practice talks, discuss experiments, or even simply think aloud has provided many insights. To the *SCAPL lab managers*, thank you for your help with all aspects of the lab. Thank you to my *interns* who have helped with data collection and data entry.

I would like to thank the *Spatial Intelligence Learning Center (SILC)*. SILC has funded my research and travel to conferences. Being at SILC has also provided me with access to a deep network of faculty, postdoctoral researchers, and graduate students across a wide range of expertise. My experiences and opportunities through SILC have been truly invaluable and deeply appreciated.

Thank you to my extremely supportive family and friends. *Marcus*, thank you for doing more than your fair share of housework and cooking. To *my family*, thank you for your support and kind words, especially *my sister* who always has my best interests at heart.

## TABLE OF CONTENTS

|  |     |
|--|-----|
| ABSTRACT.....                                      | ii  |
| ACKNOWLEDGEMENTS .....                             | iv  |
| LIST OF TABLES .....                               | vi  |
| LIST OF FIGURES .....                              | vii |
| CHAPTERS   |     |
| 1. INTRODUCTION .....                              | 1   |
| Scale Representation .....                         | 1   |
| Analogical Reasoning .....                         | 12  |
| Overview of studies .....                          | 16  |
| 2. STUDY 1 .....                                   | 23  |
| Aims .....   | 23  |
| Methods .....                                      | 27  |
| Results .....                                      | 37  |
| Discussion .....                                   | 50  |
| 3. STUDY 2 .....                                   | 61  |
| Aims .....   | 61  |
| Methods .....                                      | 63  |
| Results .....                                      | 69  |
| Discussion .....                                   | 74  |
| 4. GENERAL DISCUSSION .....                        | 82  |
| REFERENCES CITED .....                             | 91  |
| APPENDIX A: ASSESSMENT OF TEMPORAL MAGNITUDE ..... | 103 |
| APPENDIX B: ASSESSMENT OF SPATIAL MAGNITUDE.....   | 111 |

## List of Tables

1. Table 1. List of Temporal and Spatial Scales, including category names and magnitude information.....30
2. Table 2. Mean error (mm) by condition for content-specific number line estimations.....39
3. Table 3. Demographics by class and condition.....64

## List of Figures

|   |    |
|---|----|
| 1. Figure 1. Example of 3 timelines in the hierarchical alignment intervention.....   | 15 |
| 2. Figure 2. Example of a temporal and spatial number line at the thousands scale in the hierarchical condition.....        | 29 |
| 3. Figure 3. Example of conventional puzzle.....  | 32 |
| 4. Figure 4. Average error (mm) across middle events/objects, respectively.....   | 39 |
| 5. Figure 5. Response Patterns of Temporal Magnitude Estimation.....  | 42 |
| 6. Figure 6. Response Patterns of Spatial Magnitude Estimations.....  | 43 |
| 7. Figure 7. Response Patterns of Spatial Magnitude Estimations.....  | 46 |
| 8. Figure 8. Estimated location on number line by correct location for magnitudes on the million and billion scales.....    | 47 |
| 9. Figure 9. Introductory slide presented in normal class instruction.....  | 65 |
| 10. Figure 10. Example of a clicker slide.....  | 67 |
| 11. Figure 11. Example of slide containing corrective feedback.....   | 67 |
| 12. Figure 12. Comparison of control and intervention performance on exams by teacher.....                                  | 73 |
| 13. Figure 13. Comparison of control and intervention variance from correct answer on line estimation tasks by teacher..... | 73 |

## CHAPTER 1

### INTRODUCTION

Having a strong understanding of size and scale, and the relationships between scales, is essential for being a scientifically literate consumer of information (Tretter, Jones, Andre, Negishi, & Minogue, 2006a). Many fundamental scientific concepts and discoveries are at extreme scales, far removed from human experience. For example, the Geologic Time Scale, the discovery of the atom, the size of the universe, and the rapidly developing field of nanotechnology are all based on phenomena occurring at scales that cannot be directly perceived. Furthermore, scale is central to understanding important societal issues, such as governmental budgets and deficits, population growth, and global warming. Given this importance, it should be no surprise that both the *National Research Council Framework for K-12 Science Education* (NRC, 2011) and the Benchmarks for Science Literacy (American Association for the Advancement of Science (1993) have identified “size and scale” as fundamental and a unifying theme in science education. “Size and scale” was also identified as one of the “big ideas” at recent nanoscience and education national workshops (Swarat, Light, Park, & Drane, 2011).

Given the wide range of topics that involve size and scale, it is useful to take a moment to define size, scale, and what an understanding of each may mean. Size is defined as the actual magnitude of something (e.g., the weight of a mouse); scale is defined as a relative magnitude (e.g., a mouse’s weight in grams). Thus, scale links size (or magnitude) to a numerical representation of that size in conventionally defined units (e.g., grams, meters, liters, years, etc.). Understanding scale is multifaceted; it involves knowledge and accurate cognitive representations of size (or magnitude) (Barth &

Paladino, 2011), relevant measurement systems, proportional reasoning (Jones & Taylor, 2009), and properties of the given scale. Understanding properties of scales is particularly relevant for understanding relationships between different scales, as different scales operate under different physical laws. For example, while a mouse is able to maneuver effortlessly at its actual size, when scaled up faithfully to the size of a horse, the same mouse's legs would collapse under the mouse's own weight. Dimensions, such as volume and length, can also change disproportionately; that is, different dimensions do not change in direct proportion to each other.

The current dissertation explores how adults think and reason about temporal, spatial, and abstract (numeric) magnitudes at the million and billion scales. Thinking and reasoning about temporal, spatial, and abstract magnitudes may involve knowledge and accurate cognitive representations of abstract magnitudes at the million and billion scales, as well as measurement systems (years, miles, and numerical magnitude) and properties of measurement systems at different scales (e.g., millions of years may be thought about differently than hundreds of years or millions of miles).

Despite the importance of understanding size and scale in STEM (science, technology, engineering, and mathematics) disciplines, people consistently have trouble understanding and comparing sizes of very small or large magnitudes in these fields (e.g., Delgado, Stevens, Shin, Yunker, & Krajcik, 2007; Jones, Tretter, Taylor, & Oppewal, 2008; Libarkin, Anderson, Dahl, Beilfuss, & Boone, 2005; Tretter, et al., 2006a; Swarat, et al., 2011). Undergraduate students, even those in STEM majors, have difficulty in mastering concepts of size and scale (Drane, Swarat, Hersam, Light, & Mason, 2008). Stephen Hawkins (1978) has described size and scale as a critical barrier to learning and

higher-level understanding. While people are more accurate at ranking relative sizes, they struggle assigning, comprehending, and comparing absolute sizes, especially at extreme scales (Jones, et al., 2008; Tretter, et al., 2006a). For example, while people are fairly accurate on identifying a correct sequence of events on the Geologic Time Scale (Trend, 2001) and objects at astronomical distances (Miller & Brewer, 2010), they fail to understand the magnitude between the events (Tretter, et al., 2006a) and objects (Jones, et al., 2008), respectively.

Review of science education literature suggests that errors in magnitude estimation at large scales are not random, but, rather, systematic. In the following section, I will discuss how findings from research on the mental number line, category adjustment model, and event/object segmentation can help us understand common patterns of observed errors in estimation of large temporal and spatial magnitudes.

One common pattern of errors is the overestimation of relatively smaller magnitudes and the underestimation of relatively larger magnitudes for both temporal (Catley & Novick, 2008) and spatial (Tretter, Jones, & Minogue, 2006b) magnitude estimations. For example, a majority of high school students overestimate how long ago dinosaurs first appeared on Earth (Libarkin, Kurdziel, & Anderson, 2007; Petcovic & Ruhf, 2008; Resnick, Shipley, Newcombe, Massey, Wills, 2012), and underestimate when life first appeared on Earth (Catley & Novick, 2008). Estimations can vary by over 2 billion years in both cases.

This pattern of over and under estimation at large temporal and spatial scales is similar to patterns seen in number line estimation tasks at human scales, (e.g., Booth & Siegler, 2008; Dehaene & Marques, 2002; Dehaene, Izard, Spelke, & Pica, 2008; Opfer & Siegler, 2007; Siegler & Opfer, 2003). While most adults (e.g., Dehaene & Marques, 2002; Dehaene, et al., 2008) and children (e.g., Booth & Siegler, 2008) are able to use a proportional linear number line to make estimations for familiar numbers, they fail to do so with larger, unfamiliar numbers. Rather, similar to estimations of large temporal and spatial magnitudes, a pattern of over and under estimation emerges. People's estimations of magnitude increase in variation as a function of the magnitude of the judgment (Dehaene, 2003); as the value to be estimated increases in magnitude, the variation in accuracy also increases. Familiarity-based accounts of magnitude representation suggest that as people become familiar with more magnitudes, they also develop a more linear representation of those magnitudes (Barth & Paladino, 2011). For example, second-graders possess a linear representation of numbers up to one hundred, fourth-graders up to one thousand, and sixth-graders up to 100,000 (Opfer & Siegler, 2007; Thompson & Opfer, 2010; Siegler & Booth, 2004). While adults are fairly accurate making number line estimations through similar scales (Thompson & Opfer, 2010), compressive effects are visible in their response times (Dehaene, et al., 2008): people respond more slowly when making parity and size (bigger/smaller) judgments about numbers closer together compared to making judgments about numbers farther apart (distance effect), and respond more slowly when making judgments about larger numbers compared to smaller numbers (size effect).

Activation of unfamiliar magnitude representations (e.g., very large amounts of time or distances) will be less automatic than familiar values (Kadosh & Walsh, 2009). For example, people possess a weaker association between magnitude and number words for larger quantities than for smaller, more familiar quantities (Sullivan & Barner, 2010).

These findings from number line estimation tasks suggest that people represent magnitudes along a compressed mental number line (see Barth & Paladino (2011) and Opfer, Siegler, & Young (2011) for a discussion on mental models of magnitude representation). Compression refers to the representation of the distribution of magnitudes along a discrete mental number line; relatively smaller magnitudes are represented using a greater proportion of the mental number line, and the remaining (relatively larger) magnitudes are compressed into a smaller proportion of the mental number line. An explanation for why people may have difficulty estimating magnitudes at large, unfamiliar (temporal, spatial, and abstract) scales is that as magnitudes become larger on a compressed number line, they will also become less discriminable.

Another pattern of temporal and spatial magnitude estimations is the characterization of size and scale information into discrete categories. Experts formalize a categorical organization of continuous scale information (e.g., scientists divide Earth's continuous history into discrete temporal units comprising the Geologic Time Scale). People may also naturally use categories when making estimations. Huttenlocher and colleagues (1988) suggest an adaptive Bayesian model of recall of information that includes 1D, 2D, and 3D magnitudes, such as size, location, distance, and duration: the category adjustment model (CAM). The CAM suggests magnitude information is stored as a hierarchical combination of metric and categorical information. The CAM predicts

recall patterns on a range of dimensions (e.g., fatness of fish, grayness of squares, and lengths of lines (Huttenlocher, Hedges, & Vevea, 2000), events (Huttenlocher, et al., 1988), and even social dimensions such as perception of facial expressions (Roberson, Damjanovic, & Pilling, 2007) and judgments of gender and ethnicity (Huart, Corneille, & Becquart, 2005)).

There is limited research examining the CAM's predictive capability for a given dimension (such as temporal and spatial scales) across different scales (such as from human scales through to scales outside of human perception). Science education research has identified conceptual categories for spatial and temporal scales outside of human perception (e.g., Swarat, et al., 2011; Trend, 2001; Tretter, et al., 2006a); suggesting people may conceptualize magnitude information at relatively small and large temporal and spatial scales using a combination of metric and categorical information. Resnick, et al. (2012) experimentally assessed the role of categories in estimations of large temporal magnitudes. Students who were provided with salient internal structure of magnitude relations within hierarchically organized event boundaries were more likely to develop a linear representation of events on the Geologic Time Scale, compared to those who received the same information about the events without the structured magnitude information. This finding is consistent with the CAM because the provided hierarchically organized category boundaries may have made magnitude relations across scales salient, resulting in a more linear representation. Thus, this finding also suggests the use of hierarchically organized category boundaries in the representation of events and objects at larger scales.

The CAM makes specific predictions about errors in estimation, which may be useful in explaining the types of errors observed in large temporal and spatial estimations. According to the CAM, magnitude information is stored in hierarchically nested categories, and a person retrieves needed information at the level required by the question as well as the category boundaries of any associated higher-level categories (Huttenlocher, et al., 2000; Huttenlocher, et al., 1988). For example, remembering that dinosaurs first appeared in the Triassic Period implicitly contains information that dinosaurs also first appeared during the Mesozoic Era (the Mesozoic Era is comprised of the Cretaceous, Jurassic, and Triassic periods). In the absence of exact information, however, people use category boundaries of other events/objects to help make estimations. Variation in estimation, therefore, occurs because of imprecision of category boundaries (Shipley & Zacks, 2008; Zacks & Tversky, 2001). As people use event/object boundaries to help make estimations, the more imprecise or the larger the gap between boundaries, the more variation one could expect to find (e.g., Huttenlocher, et al., 1988; Shipley & Zacks, 2008). Without information at a given level, then estimations must default to a higher-level. For example, if a student cannot recall which period dinosaurs first appeared, but can recall it happened in the Mesozoic Era, their estimation of a date will range 180 million years from the Triassic period to the Cretaceous period.

While experts working with extreme scales are characterized as having a “detailed, secure, sophisticated, and well developed” mental framework of scale information, novices’ mental frameworks are found to be “scant, insecure, and nebulous” (Trend, 2000). For example, in-service science teachers represent the roughly 14 billion years of geologic events related to Earth’s history (starting with the Big Bang through to

present day) as only three conceptual categories: extremely ancient, moderately ancient, and less ancient (Trend, 2000). Thus, that novices' estimation at extreme scales (where they have few categories) may err up to five orders of magnitude (e.g., Catley & Novick, 2008) is predicted by the CAM.

Students are more accurate estimating when geologic events occurred from the Phanerozoic Eon, where they possess more categories, compared with estimations of events from the Precambrian, where they possess fewer categories (Libarkin, et al., 2005). Increased accuracy as a function of increased number of categories may be explained by category boundaries being perceptually salient. At points of unpredictability, humans are more likely to attend to information to permit more accurate future predictions (Shipley & Zacks, 2008). Event and object boundaries are defined by change (Shipley & Zacks, 2008). Subsequently, people tend to remember event/object boundaries by attending to them (Speer, Zacks, & Reynolds, 2007), and recall those events/objects at boundaries more clearly than those in between (Zacks & Tversky, 2001). Thus, students may be more accurate with estimations of events in the Phanerozoic relative to estimations of events in the Precambrian because they have more categories in the Phanerozoic. Further, providing salient category boundaries within the Geologic Time Scale fosters more accurate event estimation (Resnick, et al., 2012).

Where people hold relatively few conceptual categories for extreme scales, they also severely underestimate the magnitudes (e.g. Libarkin, et al., 2005; Swarat, et al., 2011). This may be due to subjective experience of magnitude being influenced by the number of boundaries (e.g., events/objects) the person can recall. The more category boundaries a person can recall the greater the subjective magnitude, and the converse for

the recollection of a smaller number of category boundaries (Block, 1990). Thus, there will also be a bias to allocate larger magnitudes for those regions populated with relatively more events/objects and allocate smaller magnitudes for those regions populated with relatively fewer events/objects. The more organizational structure a person has for the material the better their memory is for recall (Mandler, 1967). Where people have more conceptual categories (e.g., human scale) they are more accurate when making judgments relative to other scales (Jones, et al., 2008).

Specific to temporal estimations, a systematic bias called ‘forward telescoping’ occurs at both human (Neter & Waksberg, 1964) and geologic scales (Hofstadter, 1985; Trend, 1998, 2000, 2001; Catley & Novick, 2008). Forward telescoping refers to the tendency for people to recall events occurring more recently than they actually occurred (Huttenlocher, et al., 1988; Neter & Waksberg, 1964). Forward telescoping is also a common error people make regarding geologic events (Hofstadter, 1985; Trend, 1998, 2000, 2001; Catley & Novick, 2008). Zacks and Tversky (2001) argued that forward telescoping may occur because of an inherent asymmetry in our experience with the passage of time. A decrease in the accuracy of memory over time may result in the reliance on larger event boundaries (Huttenlocher, et al., 1988). Since there will therefore be more error in the placement of events as you move backward in time, and less error in backward dating as you move closer to the present, an aggregate forward bias of reported dates is expected by this model (Huttenlocher, et al., 1988). Subsequently, there would also be the tendency to recall events toward the center of temporally larger units (Huttenlocher, et al., 1988). If an event occurs closer to an event boundary, the increased attention to boundaries may bias the recall of that event closer to the boundary

(Huttenlocher, et al., 1988). Thus, when recalling events on the Geologic Time Scale, without salient boundaries to help guide estimations, duration of events can also be underestimated, biasing estimations towards present day. A bias towards present day is inconsistent with the overestimation of recent geologic events, like appearance of dinosaurs (Libarkin, et al., 2005). However, there may be multiple processes acting in concert that influence recalled time.

There may be other skills and strategies involved in developing or approximating a linear representation of scale. General mathematical skills (such as measurement, estimation, perspective, and proportional reasoning) and content specific knowledge (such as temperature, time, volume, and mass) have been identified in understanding scale (Jones & Taylor, 2009). The Benchmarks for Science Literacy state that by the end of eighth grade students should know “properties of systems that depend on volume, such as capacity and weight, change out of proportion to properties that depend on area, such as strength or surface processes” (AAAS, 1993, p. 278). Proportional reasoning is correlated with students’ ability to order objects and assign correct sizes to objects (Jones, et al., 2007). A particularly important aspect of proportional reasoning related to understanding size and scale is the ability to conceptualize a new unit from existing units (unitizing), and then use that new unit to make comparisons or calculations (Lamon, 1994). Experts, who use size and scale information as part of their profession, are proficient at unitizing (Tretter, et al., 2006a). For example, scientists have developed the ‘light year’ unit to better describe galactic distances.

The preceding section described the types of errors observed in large temporal and spatial estimations, and how these errors may be explained by research findings on

the mental number line, category adjustment model, and event/object segmentation.

While many people have difficulty understanding magnitudes at scales outside human perception (e.g., Delgado, et al., 2007; Jones, et al., 2008; Libarkin, et al., 2005; Tretter, et al., 2006a; Swarat, et al., 2011), experts who use size and scale in their profession often report possessing an accurate, linear representation of scale ranging from a very small (e.g., atomic) through to very large (e.g., galactic) scales (Jones and Taylor, 2009; Tretter, et al., 2006a, 2006b). One approach to identifying successful strategies in understanding scale is to examine strategies used by these experts. A strategy reported by experts is to use anchor points relevant to their field as referents to help estimate unknown or less used magnitudes (Jones & Taylor, 2009). Commonly used objects are associated with specific sizes (e.g., the red blood cell is  $7\mu\text{m}$  across), and then the object itself becomes the standard for other measurements. Experts also report visualizing the scaled world in which they are operating (Jones and Taylor, 2009; Tretter, et al., 2006a, 2006b). They note not moving continuously from one scale to the next, but, rather, making a discontinuous leap to the scale they need to think about, maintaining abstract connections between different scales using mathematics (Tretter, et al., 2006b).

The most frequently cited strategy among experts in understanding extreme scales is analogical reasoning (Jones and Taylor, 2009). Experts utilize models and analogy as a way to physically experience different scales (Tretter, et al., 2006b). Analogy and visual displays are also the most commonly used pedagogical practices (Libarkin, et al., 2007). Thus, the current studies examine how analogical principles can be used to learn about large temporal and spatial scales.

## Analogical Reasoning

Analogy refers to relating two different concepts (a base and a target concept) based on shared features. Analogies are powerful classroom tools because they allow for understanding of an unfamiliar concept through familiar concepts. The football field is often used as an analogy for extreme scale representations, such as for geologic time (Wheeling Jesuit University, 2004) and the solar system (Jones, Taylor, & Broadwell, 2009). For example, one end line of the football field would represent when Earth forms, the other end line would represent present day, and students would learn on this scale humans appeared within the first yard line closest to present day. While being able to kinesthetically experience a size or scale correlates with comprehension of that scale (Tretter, et al., 2006b), educators must rely on representations, metaphors, and analogies to teach concepts at extreme scales (e.g. outside of human scale ranges) because very small and very large scales cannot be directly experienced (Jones, et al., 2009).

Spatial analogies are particularly well suited for thinking and learning about temporal and spatial scale information for a number of reasons. Extreme temporal and spatial scales cannot be directly experienced, and, therefore, mapping these scales onto another dimension (such as a space that can be experienced) may approximate experience with the extreme scale (Jones & Taylor, 2009; Tretter, et al., 2006a, 2006b). Temporal and spatial magnitudes are also cognitively represented in similar ways (e.g., Gentner, 2001; Lakoff & Johnson, 1980; Walsh, 2003) (although the precise nature of the relationship is debated (Boroditsky, 2001; Gentner, 2001)). Further, there is a correspondence between time and space in the external environment. We can literally see a mapping of time onto space, as newer rocks form on top of older rocks.

Unfortunately, there are barriers to using analogy as a learning tool. It is possible for analogies to fail to bring about conceptual change (Brown & Salter, 2010; Duit, 1991). Analogies can even mislead students' understanding of a concept, making misconceptions hard to identify and resolve (Brown & Salter, 2010; Duit, 1991). This may occur due to there being salient features that are not functional to the analogy. Not all features are equally relevant; some attributes may be more accessible than others (Gentner, 1983). Salient surface features may influence the analogy made (Brown & Salter, 2010; Gentner, 1983). For example, biology classrooms often draw an analogy between water pipes and blood vessels to demonstrate their functionality of carrying liquid (Brown & Salter, 2010). However, the salient property of pipes being rigid can create misconceptions regarding the elasticity of blood vessels, which is a critical attribute of how blood vessels function. Additionally, similar misconceptions are fostered due to differing surface features between water and blood.

Barriers to alignment may be why students continue to struggle to understand large magnitudes despite a wide range of pedagogical approaches. One approach to teaching large temporal and spatial magnitudes is to align the extreme scale with a familiar scale (e.g., Clary & Wandersee, 2009). However, given the difference in magnitude between familiar (e.g., human) and unfamiliar (e.g., galactic) scales, it may be difficult to align the shared features of both scales: the base concept and target concept are too different. Another analogical approach is mapping the extreme magnitude onto a structure (e.g., a roll of toilet paper, a cross-country road trip, and a football field (Clary & Wandersee, 2009) are often used as analogies for the Geologic Time Scale); however, while the base concept may be familiar, the magnitude of the base concept may not be

(e.g., how long is a roll of toilet paper?). There may also be psychological barriers to alignment based on pre-existing spatial or functional characteristics. For example, in a cross-country analogy, state divisions of varying lengths may bias magnitude recall (e.g., Friedman & Brown, 2000; Stevens & Coupe, 1978). Practical constraints may also serve as a barrier to alignment. For example, if aligning extreme magnitude to a roll of toilet paper, the physical size of the classroom may necessitate bending the roll of toilet paper, making it more difficult to accurately represent the magnitude of the base concept and subsequent magnitude relations between scales.

One way to overcome barriers to alignment between a base concept and a target concept is through hierarchical alignment (Resnick, et al., 2010), which is based on the progressive alignment approach to analogical reasoning (Kotovsky & Gentner, 1996; Thompson & Opfer, 2010). The progressive alignment model approach to analogical reasoning advocates the comparison of two highly similar items. The more commonalities that exist between these items, and if these commonalities are highlighted, the more salient corresponding relations will be. Thus, comparing two very similar items will help extend the analogy to unfamiliar items (Gentner & Namy, 2006). Furthermore, the act of performing comparisons may change original mental representations, increasing the uniformity between the two representations. The process of alignment may make higher-order relational similarities more salient. Recognition of higher-order relational commonalities may promote making the same subsequent higher-order connections with unfamiliar items (Kotovsky and Gentner, 1996). The progressive alignment of scales may alleviate the conceptual dissimilarity between human scales and extreme scales by providing more structural alignment across smaller increases of scale.

The hierarchical alignment approach to analogical reasoning progressively aligns concepts, and also hierarchically organizes each base concept within the new target concept (Resnick, et al., 2010). For example, in a successful intervention using hierarchical alignment to foster a linear representation of the Geologic Time Scale, students made ten separate time lines, each 1 meter long: a personal time line, average human lifespan, American history, recorded history, human evolution, Cenozoic Period, Phanerozoic Eon, Proterozoic Eon, Archean Eon, and Hadean Eon. For each timeline, students located where all previous timelines would begin on the current timeline (see Figure 1). For example, for the recorded history time line, recorded history was represented as one meter. On the following time line, human evolution, human evolution was represented as one meter and recorded history was represented as less than one centimeter. This hierarchical organization highlights how each temporal scale is proportionally related to one another, helping to populate each scale with boundary information by providing internal structure of magnitude relations within event boundaries.

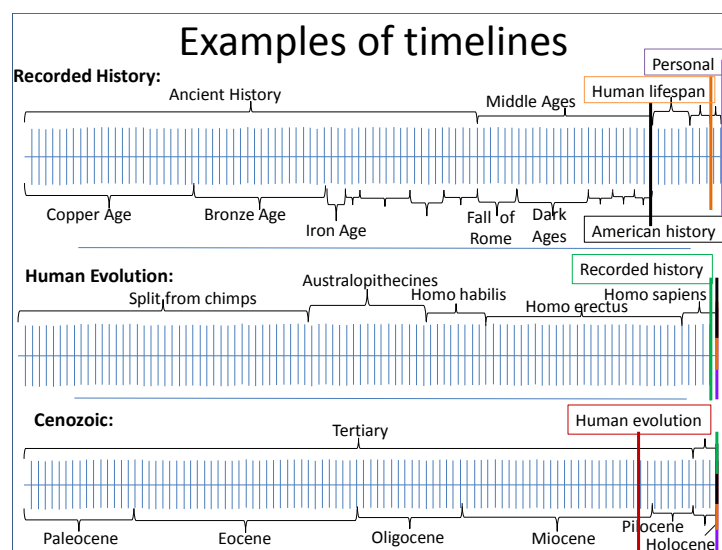


Figure 1. Example of 3 timelines in the hierarchical alignment intervention

## Overview of Studies

The current studies are rooted in the science of learning; that is, they aim to leverage the way people naturally think and reason into effective pedagogy. While looking to cognitive science to answer the question of why people have trouble understanding scale information is not a new idea (e.g., Dodick & Orion, 2003; Trend, 2001), prior work has not considered the role of abstract magnitude representation in understanding scientific phenomenon at large scales, the use of principles of analogical reasoning in developing analogies for scale information, scientific assessment of analogies for scale information, and leveraging such relevant cognitive science findings into learning tools. In the following section I will discuss how the current studies aim to fill these gaps.

There are a number of studies that acknowledge magnitude representation as a potential contributing factor in understanding scale information (e.g., Catley & Novick, 2009; Semken, et al., 2009; Trend, 1998, 2000, 2001). However, a majority of these studies do not actually assess the role of magnitude representation in understanding scale information; rather, they examine ability to sequence stimuli in the correct order (Cheek, 2012). Two studies that do examine magnitude representation (e.g., Cheek, 2012; Lee, Liu, Price, & Kendall, 2010) assess ability to make estimations only within a single content domain (e.g., geologic time).

Studies that examine only one domain are unable to differentiate between the contributions of variables such as knowledge of that domain (e.g., if you have never heard of banded iron formations, how would you be able to estimate when they appeared?), cognitive representations of that domain (e.g., temporal reasoning may be different than reasoning about other magnitudes), as well as cognitive representations of abstract magnitude (e.g., having a linear representation of relevant abstract (numeric) magnitudes). There is currently only one study that directly examines the role of magnitude representation in understanding scale information (Resnick & Shipley, in prep).

Resnick and Shipley (in prep) investigate the role of temporal and abstract magnitude representation in understanding geologic time. Students enrolled in an introductory geoscience course learned about geologic time by either completing a shortened stratigraphy lab plus an activity that used analogical principles of hierarchical and progressive alignment (intervention group) or by completing a stratigraphy lab of equal duration (control group). In the stratigraphy lab, students identified the ages of layers of rock in images and diagrams of rock exposures. Thus, in both conditions students practiced mapping time onto space. An item from the Geologic Concept Inventory, which is a valid and reliable instrument measuring a range of geoscience concept knowledge (Libarkin, et al., 2005), was used to measure understanding of geologic time. For this item, students were presented with five time lines populated with four events, and asked to choose the time line with events in the correct proportional locations. Students also completed four number line estimation tasks, as a measure of large abstract magnitude representation. In the line estimation task, students are asked to

estimate where specific magnitudes are positioned on a horizontal number line. The number line represented 4.6 billion years: the left flank of the number line was labeled as '0 years ago' and the right flank was labeled as '4.6 billion years ago'. The magnitudes used were equivalent to the number of years ago the events occurred on the Geologic Concept Inventory item (6 million, 65 million, 230 million, and 3.5 billion).

The intervention group demonstrated a more accurate sense of the relative durations of geological events and a reduction in the magnitude of temporal location errors relative to the control group. This suggests that the hierarchical and progressive alignment of geologic time is an effective way to reduce magnitude-based errors in understanding geologic time. There was also an overall difference in performance between temporal and abstract magnitude estimations; with a majority of students possessing an incorrect temporal representation of geologic events, and performing near ceiling for abstract magnitude. This suggests that temporal and abstract magnitudes are dissociable. Differences in abstract magnitude representation between the intervention and the control groups were only visible when students were asked to estimate 230 million on a 4.6 billion number line. Given the near ceiling performance on the three other abstract magnitude estimations (6 million, 65 million, and 3.5 billion), the authors hypothesize the other three magnitudes are close to salient anchor points, allowing for more accurate estimation. Estimations of 6 million and 65 million are located within one mm of the left flank (0 years ago). The estimation of 3.5 billion is at 7.6cm on a 10 cm scale, which is within one mm of a quarter boundary ( $3/4$ ), which may also be salient.

A main aim of the current studies is to replicate and extend the findings from the hierarchical alignment study (Resnick & Shipley, in prep). The current studies will

examine if the hierarchical alignment model, and specific components of the hierarchical alignment model, are effective in fostering a linear representation of temporal, spatial, and abstract magnitudes. In doing so, the current studies will also explore how people normally represent magnitudes in these different domains. By assessing temporal, spatial, and abstract magnitude representation, the current studies will be poised to investigate if magnitudes are represented similarly across domains. The current studies will also be able to examine if learning a linear representation of magnitude in one domain (e.g., temporal) transfers to other domains (e.g., spatial), or if having a linear representation of abstract magnitude transfers to content-specific domains (e.g., temporal).

To measure magnitude representation, the current studies expand upon the measures used by Resnick & Shipley (in prep). Line estimation tasks are used for assessing temporal, spatial, and abstract magnitudes in study 1, and for assessing abstract magnitudes in study 2. Line estimations enable researchers to assess patterns in magnitude representation. Student performance on the abstract magnitude task used by Resnick & Shipley (in prep) was near ceiling, with group differences appearing only for the estimation of 230 million on a 4.6 billion number line. The authors hypothesize the other three magnitudes (6 million, 65 million, and 3.5 billion) are close to salient anchor points, allowing for more accurate estimation. The current studies will use abstract magnitude estimations hypothesized to be farther away from salient anchors points (e.g., 542 million on a 4.6 billion scale). Line estimations will also be at the million and billion scale, to examine differences in representation of magnitudes at either scale.

Temporal and spatial line estimation tasks that ask participants to estimate a number of years or miles (e.g., “where would 230 million years ago be located on this time line?”) may possess very similar surface characteristics with abstract (numeric) magnitude estimations. Specifically, number of years may be interpreted as years or as an abstract magnitude (e.g., “230 million”). Thus, study 1 will frame abstract magnitude as temporal or spatial, and compare performances on both types of line estimations. Study 2 will include temporally framed abstract magnitude estimations. Framing abstract magnitude estimations as temporal or spatial will also help keep the assessment measures consistent with the learning tasks. Consistency is particularly important with the classroom-based study. Temporal and spatial estimation will use non-numerical flanks (e.g., “Present day”, “Earth’s surface”, etc).

Assessing magnitude representation at two different large scales serves as an initial exploration to magnitude representation at different scales. There is currently limited research that directly examines the representation of magnitudes required for understanding scientific phenomenon, such as geologic time and astronomical distances. Studies examining unfamiliar magnitude representation generally focus on younger children (e.g., Booth & Siegler, 2008; Ebersbach, et al., 2008; Izard & Dehaene, 2008; Thompson & Opfer, 2010), with magnitudes ranging up to 10,000 (Thompson & Opfer, 2010). While there are a few studies that look at magnitude representation in the adult population for much larger scales (e.g., billions) (Landy, Silbert, & Goldin, 2012; Resnick & Shipley, in prep), and unfamiliar magnitudes (conventional (million, billion) and fictitious magnitudes (e.g., zillion) (Rips, 2012), more research is needed to truly characterize adult representation of extreme magnitudes.

The current studies aim to extend the findings from Resnick, et al. (2012) by examining the effectiveness of the hierarchical alignment model in other domains (i.e., spatial magnitudes) (study 1), as well as assessing the necessity of two components of the hierarchical alignment model. The hierarchical alignment model is comprised of a number of principles (e.g., progressive alignment, hierarchical alignment, structural alignment, and multiple opportunities to make relevant analogies). The current studies will assess the importance of two of these principles (progressive and hierarchical alignment). In study 1, the benefit of hierarchical alignment is assessed by comparing conditions using hierarchical alignment and progressive alignment with conditions that just use progressive alignment. In study 2, students complete an activity based on all the same principles as study 1, except, scale information is not hierarchical or progressively aligned. Thus, study 2 examines the necessity of hierarchical alignment and progressive alignment. By identifying which principles of the hierarchical alignment model are necessary, an activity can be developed that fits more naturally into an already full curriculum (i.e., an activity that requires less time to complete).

More progress is needed in the translation of cognitive science research into learning tools for teaching extreme magnitudes. With the exception of Thompson and Opfer (2010) and Resnick, et al. (2012), existing research on domain-specific magnitudes (e.g., temporal and spatial magnitudes), which take into account cognitive science research, can generally be categorized as possessing one of the two following aims: identify that students have difficulty (e.g., Delgado, et al., 2007; Trend, 1998) or develop assessments of student ability/knowledge (e.g., Dodick & Orion, 2006; Libarkin, et al., 2005). Further, given the prevalence and importance of analogies in teaching extreme

scales (Libarkin, et al., 2007), it is surprising the scarcity of studies that leverage principles of analogical reasoning when developing an analogy for scale information, as well as lack of scientific testing in the assessment of an analogy's effectiveness in teaching scale information. A majority of published analogies are teacher recommendations, with no scientific testing of their effectiveness (e.g., Clary & Wandersee, 2009; Richardson, 2000; Wheeling Jesuit University, 2004). There are some studies that aim to scientifically assess the effectiveness of an analogy (e.g., Hermann & Lewis, 2004; Petcovic & Ruhf, 2008; Semken, et al., 2007); however, these studies do not utilize cognitive science research on principles of analogical reasoning in the development or the assessment of the analogies. There are currently two studies that utilize principles of analogical reasoning in the development of an analogy for unfamiliar magnitudes, and scientifically assess the analogy's effectiveness: Thompson & Opfer's (2010) application of progressive alignment (Kotovsky & Gentner, 1996) to learning to reason about unfamiliar abstract magnitudes for young children, and Resnick, et al.'s (2012) application of hierarchical alignment to learning to reason about large temporal magnitudes for undergraduates. The current study aims to use the science of learning to develop an effective learning tool for fostering linear representations of magnitudes across different dimensions, and thus also develop our understanding of how people learn to reason about unfamiliar magnitudes.

## CHAPTER 2

### TRANSFER STUDY

#### Aims

The current study is multifaceted. An overarching aim is to examine if the hierarchical alignment activity (Resnick & Shipley, in prep) is effective in developing a linear representation of temporal, spatial, and abstract magnitudes. Success of the hierarchical alignment activity would have both educational and theoretical implications. In educational contexts, the hierarchical alignment activity may be a useful tool for educators teaching scientific phenomenon at extreme scales (e.g., geologic time). Theoretical implications include informing the Category Adjustment Model (CAM), magnitude representation across scales and domains, and principles of analogical reasoning. These theoretical implications are detailed below coupled with specific aims of the current study.

The hierarchical alignment activity is based on the hypothesis that magnitudes outside of human perception are represented the same way as magnitudes at human scales. Briefly, magnitude information at human scales may be stored as a hierarchical combination of metric and categorical information, and category boundaries are used to make estimations in lieu of exact metric recall (e.g., Crawford, Huttenlocher, & Hedges, 2006; Huttenlocher, et al., 1988; Newcombe, Huttenlocher, Sandberg, Lie, & Johnson, 1999; Zacks & Shipley, 2008) (see Chapter One for more detail). The hierarchical alignment activity provides participants with salient category boundary information by providing internal structure of magnitude relations across scales. Specifically, the hierarchical alignment activity provides salient boundary relations by allowing

participant to practice seeing the relative relationships between magnitudes at different scales. Thus, success of the hierarchical alignment activity for temporal, spatial, and abstract (numeric) magnitudes would provide evidence that magnitude information across domains and scales are represented and reasoned about in similar ways. The CAM has been a useful model for predicting patterns of recall for temporal and spatial magnitudes at human scales. Thus, if the hierarchical alignment activity is successful at developing linear representations of temporal, spatial, and abstract magnitudes outside of human perception, the CAM might be used in predicting patterns of recall at scales outside human perception as well.

By measuring the effectiveness of the hierarchical alignment activity, the current study also explores how people normally represent temporal, spatial, and abstract magnitudes. Differences (and similarities) in performance will inform the topical debate of how magnitude representations are represented in the brain. A generalized mapping of more/less relations across dimensions (such as time, space, number, and size) has been suggested (Walsh, 2003). Such mappings have been observed in preverbal infants (Brannon & Roitman, 2003; Lourenco & Longo, 2010) and in non-human animals (Brannon & Roitman, 2003). However, inconsistent findings in neuroscience literature (Agrillo, Ranpura, & Butterworth, 2010) and asymmetrical relationships in cross-dimensional interference paradigms (Agrillo, et al., 2010; Casasanto & Boroditsky, 2008) is inconsistent with a model of a generalized system of magnitude. If the hierarchical alignment activity is successful in developing a linear representation of temporal, spatial, and abstract magnitudes; and if performances on temporal, spatial, and abstract magnitudes are similar; it would suggest magnitude information across domains are

represented and reasoned about in similar ways. The current study also investigates specific relationships between representations of temporal, spatial, and abstract magnitudes. For example, the role of having a linear representation of large magnitudes (either domain-specific or abstract) on understanding and estimation of relevant scientific phenomenon is examined. Additionally, transfer is assessed between learning a linear representation of magnitude in one domain (e.g., temporal) to another domain (e.g., spatial).

Recall of temporal categories and individual temporal magnitudes is examined. An individual temporal magnitude is defined by some amount of time (e.g., 100 million years ago). A temporal category is defined by grouping a number of years together by a common attribute (e.g., the Cretaceous Period was named for extensive chalk deposits in Europe, and lasted from 145 million years ago to 65 million years ago). Participants may have better recall for temporal categories versus temporal magnitudes (or vice versa). Differences in memory span have been identified for a range of things; such as digits, colors, and geometric designs (Crannell & Parrish, 1957). People appear to be able to recall more Arabic numerals than number words (e.g., Chincotta and Underwood, 1997), a phenomenon referred to as the numeral advantage effect. Recall of temporal categories and temporal magnitudes may also be related; categorical information may aid in the recall of individual temporal magnitudes through 'chunking'. Chunking refers to people being able to recall more items or numbers when they group (or chunk) those items or numbers together in a meaningful way (Miller, 1956).

The current study assesses analogical reasoning principles for learning magnitudes outside human perception. The hierarchical alignment activity is comprised of a number of analogical reasoning principles, based on hierarchical (Resnick & Shipley, in prep) and progressive (Kotovskiy and Gentner, 1996; Thompson & Opfer, 2010) alignment. The current study contrasts the hierarchical alignment activity with another activity comprised of principles of progressive alignment. Thus, the additive benefit of the hierarchical alignment of scale information on learning is assessed.

To address these specific aims, participants are presented with information about temporal and spatial magnitude either hierarchically or conventionally (hierarchical and conventional procedures are described in detail below). Temporal magnitudes are always presented first and spatial magnitudes second across conditions. A fixed order presentation reduces the number of conditions, thus making the length of time the study takes to complete manageable, while still addressing identified research questions (e.g., is the hierarchical alignment activity effective, are there additional benefits of hierarchical alignment, how are magnitudes across domains naturally represented, does learning about temporal magnitudes transfer to learning about spatial magnitudes). Findings from the current study will inform future directions and research questions.

## Methods

### Participants

Eighty individuals (60 female), ages 18-29 years old ( $\mu=21.41$ ,  $\sigma=3.98$ ), participated in this experiment. Participants were recruited from an undergraduate psychology experiment pool at a large urban American university in exchange for course credit. The study sample was comprised of 63% participants identifying as “Caucasian”, 18% identifying as “African-American”, and 19% identifying with another ethnicity (including “Haitian”, “Jamaican”, “Trinidadian”, “Vietnamese”, “Asian”, and bi-racial). Education levels include 19% Freshman, 18% Sophomore, 13% Junior, 23% Senior, and 27% identified themselves as an “undergraduate student”. Fifty-six percent of students had previously taken a geoscience course in either high school or college.

### Materials

**Hierarchical Design** In the hierarchical alignment condition, participants completed the same hierarchical alignment activity developed by Resnick and Shipley (in prep), which is based on the progressive alignment model (Kotovsky and Gentner, 1996; Thompson & Opfer, 2010). Participants made ten separate time lines, aligning time to a one meter space. They began by making a personal time line. A personal time line was chosen as the base concept, because participants should be familiar with their own personal history as well as mapping human temporal scales onto space (i.e., making time lines). The participants then made nine other time lines; working through different historic and geologic time lines, up to the full Geologic Time Scale (see Table 1). Each time line was chosen for use in the hierarchical alignment activity based on conventionally defined boundaries (e.g., the Archean, Proterozoic, and Cenozoic are all

divisions in the Geologic Time Scale) that differ by orders of magnitude. For each time line, students were presented with a partially completed time line (see Figure 1 and Figure 2). Participants were required to label the time line's length (in years) and locate where all previous time lines would begin on the current time line. This hierarchical organization highlights how each temporal scale is related to the others. Hierarchical organization helps to populate each scale with boundary information by providing internal structure of magnitude relations across scales. To figure out where previous time lines were located, participants were given two mathematical equations: one to determine how many years each centimeter would equal (number of years time line represents/number of centimeters (always 100)), and another equation to determine how many centimeters were needed to make up previous time lines (number of years previous time line represents/number of years each centimeter represents). Help completing math was provided as needed. After all previous time lines were located, participants were then told information about events on that time line. After the completion of each time line, the completed time line was taken away; so only one time line was visible at a time.

The current study developed an analog version of the temporal hierarchical alignment activity for spatial distances (see Table 1 and Figure 2). For the hierarchical alignment of spatial distances, participants align ten increasingly larger scales of distance to a one meter space, beginning with a familiar distance. The ten spatial scales were chosen for use in this study by aligning distances with the temporal orders of magnitude used in the temporal activity. While temporal and spatial magnitudes were equated on orders of magnitude, none of the individual magnitudes were the same. For example, the Cenozoic time line and the Mercury number line are matched on the scale of tens of

millions, but both time lines represent different magnitudes (65 million and 57 million, respectively). There are two practical reasons why the individual temporal and spatial magnitudes are not the same. For both temporal events and spatial objects, real scientific data was used; thus, it is extremely unlikely that temporal and spatial magnitude divisions would be entirely aligned. Additionally, having the exact same magnitudes for time and space may draw participant attention to the experimental objectives (e.g., thinking about magnitudes), which may change participants' normal learning behavior. The hierarchical alignment condition takes approximately 45 minutes to complete.

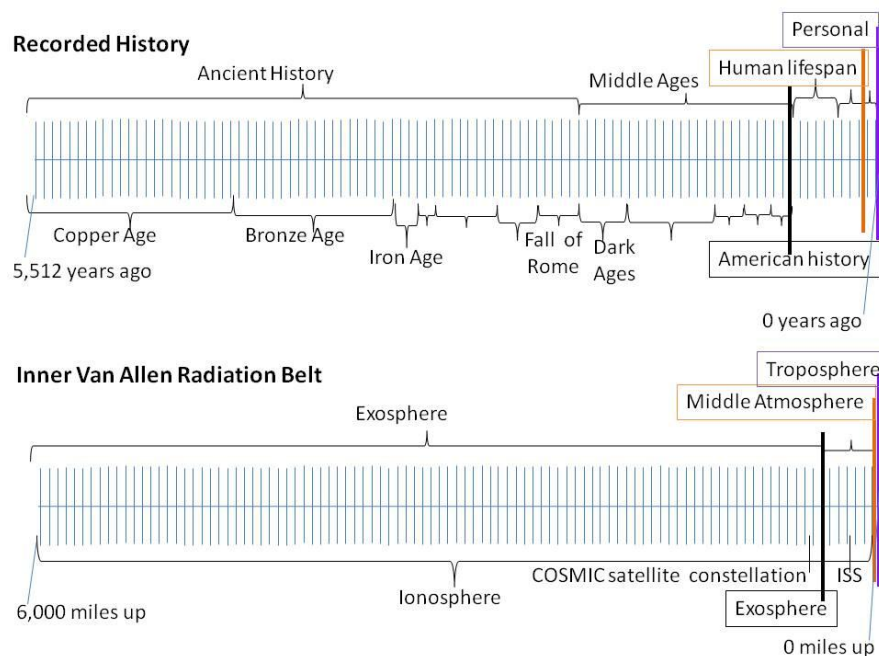


Figure 2. Example of a temporal and spatial number line at the thousands scale in the hierarchical condition. Note: the three previous temporal and spatial number lines are located relative to the current scale.

Table 1. List of Temporal and Spatial Scales, including category names and magnitude information

| Temporal Scale   | Years         | Spatial Scale                   | Miles         |
|------------------|---------------|---------------------------------|---------------|
| Personal         | 20            | Troposphere                     | 11            |
| Human Lifespan   | 75            | Middle Atmosphere               | 52            |
| American History | 519           | Exosphere                       | 400           |
| Recorded History | 5,512         | Inner Van Allen Radiation Belt  | 6,000         |
| Human Evolution  | 6,000,000     | 3753 Cruithne (quasi-satellite) | 8,450,000     |
| Cenozoic         | 65,000,000    | Mercury                         | 57,000,000    |
| Phanerozoic      | 542,000,000   | Saturn                          | 777,000,000   |
| Proterozoic      | 2,500,000,000 | Neptune                         | 2,700,000,000 |
| Archean          | 3,800,000,000 | Pluto                           | 3,580,000,000 |
| Hadean           | 4,600,000,000 | Makemake (dwarf planet)         | 4,800,000,000 |

**Conventional Design** The study sought to contrast the intervention with a realistic training program similar to one that might be used to instruct students in a classroom on these scales. Common pedagogical approaches to teaching geologic time (Libarkin, et al., 2007) and astronomical distances (Miller & Brewer, 2010) are to create spatial analogies, such as placing events/objects in the correct sequence. In order to examine the effects of hierarchical alignment, the experimental and control conditions were matched on the following properties: number of time lines, number of times participant identifies each previous scale (i.e., the first scale is identified ten times in relation to the current scale; the last scale is identified just once), progressive increase of magnitude, information provided about each event/object, and total length of time on task.

Participants completed ten separate puzzles, placing the events/objects into the correct sequence. The puzzles were made up of pieces of paper, half containing magnitude information and half with the respective category information. Participants were required to match the magnitude information with the corresponding category information for each scale, and place the scales in the correct sequence. The first puzzle represented the first temporal/spatial scale (see Table 1), with each puzzle representing an increased amount of magnitude. The tenth and final puzzle represented all of geologic time/distance to Makemake. Participants were told the same information about the events/ objects at each scale as in the hierarchical condition. After the completion of each puzzle, the puzzle was taken away; so that only one puzzle was visible at a time. The conventional condition took approximately 45 minutes to complete. Thus, the only difference between conditions was the hierarchical alignment of scale information.

One potential difference between the temporal and spatial information was identified. Participants are likely to be familiar with thinking about temporal scales extending back hundreds of years ago; learning about recent human history is common. However, participants may not have the same level of familiarity with conceptualizing the vertical nature of the spatial scales. Because it is likely people have more experience traveling parallel to Earth's surface, or 'horizontally', as opposed to traveling vertically away from Earth's surface, we used this horizontal experience as an initial introduction of the vertical scale. As a way to familiarize participants with the vertical scale, a horizontal map was presented for each of the first three scales in both the hierarchical and conventional conditions. The maps showed an eleven, fifty-two, and four-hundred mile radius extending out from the university where the study took place. To engage the

participants in grounding this scale to their personal experience, participants were asked if they had been anywhere on that radius or if they were familiar with the area. No map was provided for the remainder of the spatial scales, since these larger spatial scales are likely equally familiar to participants as the matched temporal scales.

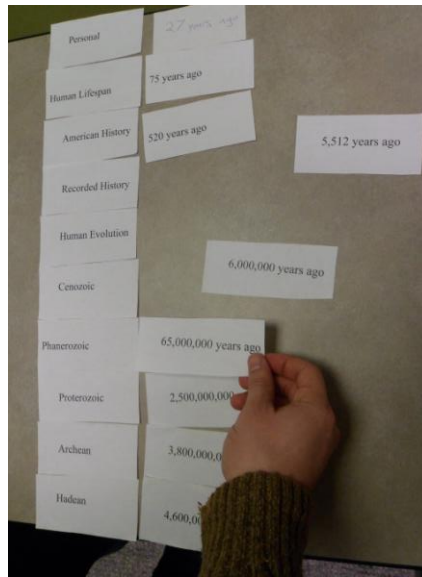


Figure 3. Example of conventional puzzle.

Note: In this example, the participant completes the last temporal puzzle. Note: there are ten scales (ten pieces of paper with magnitude information, and ten pieces of paper with category information). The participant places the pieces of paper in the correct sequence on a vertical axis.

### Procedure

All participants read an information sheet and signed a consent form. Participants completed pretest measures: a hard-copy packet containing questions (fixed order) on proportional reasoning and abstract magnitude representation. Individuals then participated in one of four conditions. In each condition, participants were presented with information about time and then distance. Information about time and distance was presented either hierarchically or conventionally (~90minutes). Thus, the four conditions include learning: 1) time hierarchically and space hierarchically (THSH), 2) time

conventionally and space hierarchically (TCSH), 3) time hierarchically and space conventionally (THSC), and 4) time conventionally and space conventionally (TCSC). There were 20 participants in each condition (80 total). Participants across conditions then completed the outcome measures. Participants first completed a hard copy packet for temporal questions, and then a hard-copy packet for spatial questions. Lastly, participants answered questions regarding demographics. The entire study took approximately 2 hours to complete.

### Measures

A series of line estimation tasks were developed to assess participants' representations of geologic time, astronomical distances, and abstract numerical magnitude (see Appendices A and B). All line estimation tasks were presented on a vertical number line 173.5mm in length.

To measure representation of events on the Geologic Time Scale, an item from the Geoscience Concept Inventory, a reliable and valid instrument measuring a range of geoscience knowledge (Libarkin, et al., 2005), was adapted as a number line task. The Geoscience Concept Inventory item presents participants with five time lines, with the following four geologic events placed in different locations: "life appears", "dinosaurs appear", "dinosaurs appear", and "humans appear". Participants are required to choose the correct linear representation, with the other four time lines representing common misconceptions (life occurred when Earth formed, humans and dinosaurs coexisted, dinosaurs appeared much earlier than they did, and all life formed at the beginning of Earth's history). While this item has been previously successful in assessing participant representations of geologic time (e.g., Libarkin, et al., 2005; Petcovic & Ruhf, 2008;

Resnick, et al., 2012), pilot testing (twenty participants; five participants in each condition) suggest this item may not be sensitive enough when assessed immediately after learning (pilot participants were at ceiling). This difference in performance may be due to laboratory vs. classroom settings. In order to more sensitively measure the variance in participants' representations immediately after learning, the Geoscience Concept Inventory item was adapted so that participants were given a blank time line (anchored by "present day" and "Earth forms"), and asked to locate the same four events as used in the Geoscience Concept Inventory item: "life appears", "dinosaurs appear", "dinosaurs disappear", "humans appear".

To measure representation of objects on an astronomical scale, an item was developed as an analog to the geologic event time line described above. Here, participants were presented with a blank number line (anchored by "Earth's surface" and "Makemake"), and asked to locate four objects on the same scale as on the event time line: "Pluto", "Mars", "Mercury", and "Cruithne".

To measure representation of abstract (numeric) magnitude (not content specific) a series of line estimation tasks were given. Participants were given a sentence stating when/where an event/object was, and then asked to locate that magnitude on the number line (e.g., "Venus is 26 million miles away from Earth. Please draw on the line provided where Venus is located."). These items were framed in terms of events and objects to match the form of the other experimental measures. These estimations are considered estimations of abstract numerical magnitude because the participants are explicitly given a magnitude to place on the number line; no recall is required. The questions provide the numerical values and ask for an estimation of the appropriate location on a spatial scale.

To assess representations of the million and billion scales, participants were asked to estimate two ‘events’ and two ‘objects’ on a 4.6 billion scale, and two ‘events’ and two ‘objects’ on a 542 million scale.

A series of multiple choice items were developed in order to assess understanding of scientific concepts involving extreme temporal (see Appendix A) and spatial scales (see Appendix B). There are twelve temporal and twelve spatial items that are direct analogs. Eight of these items require magnitude recall only (e.g., “When did dinosaurs disappear?”). To measure differences in recall of categorical and magnitude information, four items included category response options (e.g., “A. Triassic...”), and four included magnitude response options (e.g., “A. 65 million years ago...”). While participants answered the same question with both types of response options, repeated questions were presented on separate pages so participants could not compare answers. The remaining four items required magnitude recall plus an additional step of reasoning (e.g., “What is the relationship between dinosaurs disappearing and humans appearing?”). In this example, participants are required to recall when dinosaurs disappeared, when humans appeared, and compare the relative durations in between.

One of the multiple choice items for temporal information was developed by Barghaus and Porter (2010) for use with middle school students. In this item, participants use an image of the Geologic Time Scale to identify a true statement (see Appendix A). Despite being provided with precise magnitudes, the true statement (“The Proterozoic lasted much longer than the Phanerozoic”) is counterintuitive because the amount of space provided on the Geologic Time Scale is not aligned with the magnitudes. An analog version of this item was created for spatial information. The remaining items were

developed through collaboration between a cognitive psychologist and a geologist specializing in scientific phenomenon that occur at large scales. An item from the Geoscience Concept Inventory (Libarkin, et al., 2005) was included for the temporal packet only as a thirteenth multiple choice item. This item asked participants to choose the statement that best described the Earth when it first formed. There is no spatial equivalent for this question.

Participants were classified as “poor at math”, “strong at math”, or “average at math” based on their performance during participation. “Poor at math” is characterized as having difficulty completing basic mathematical tasks, such as dividing numbers by one hundred (e.g.,  $400/100=4$ ). “Strong at math” is characterized as demonstrating a mastery over more complex mathematical tasks, such as mentally dividing large numbers quickly. For example, participants who divided numbers like 3.5 billion by 46 million quickly using no paper and pencil were labeled as “strong at math”. “Average at math” is characterized as being able to complete basic mathematical tasks and having minor to no problems, and using paper and pencil with more complex mathematical tasks. Each participant was classified based on observations made while they completed the math required in the construction of the temporal and/or spatial number lines. Participants who learned both the temporal and spatial information conventionally (TCSC) were unable to be assigned a classification because they did not perform any mathematical tasks with the experimenter (i.e., the participants in this group did not construct any number lines). As the experimenter worked one-on-one with the participant, inter-coder reliability was not obtained. Thus, interpretations of any findings involving math skill should be considered as preliminary.

Participants completed two pre-test measures. Basic proportional reasoning ability was assessed using proportional reasoning measures from Park, Park, and Kwon (2010). Abstract linear magnitude was assessed using four number line estimation tasks. Participants were asked to identify where an abstract magnitude would be located (e.g., Where is 400 million on this number line). Consistent with the other number line estimation tasks, two of the number lines represented 542 million and two represented 4.6 billion. At the end of the study, demographic information (e.g., sex, ethnicity, age, etc) was obtained, including if the participant has previously taken a geoscience course.

### Results

There are four conditions: one where participants learn about temporal and spatial magnitude hierarchically (THSH), one where participants learn about temporal and spatial magnitudes conventionally (TCSC), and two where participants learn about temporal and spatial magnitudes using a combination (THSC and TCSH). Error on all number line estimations (temporal, spatial, and abstract) were calculated by taking the absolute distance (in mm) of a given response from the correct location.

*Does the hierarchical alignment activity develop a linear representation of domain specific magnitude?* The temporal and spatial line estimations were direct analogs of one another. Temporal estimations were made on a scale from “present day” (0 years ago) to when “Earth formed” (4.6 billion years ago). Spatial estimations were made on a scale from “Earth’s surface” (0 miles away) to “Makemake” (4.8 billion miles away). There were four number line estimations of temporal and spatial magnitudes. Event/object one is the closest to the zero flank, on the scale of millions. Event/object two is on the scale of tens of millions, and event/object three is on the scale of hundreds

of millions. Event/object four is closest to the end flank, on the scale of billions. See Table 2 for mean error for individual temporal and spatial estimations by condition. See Figure 5 and Figure 6 for images of correct responses and common errors for temporal and spatial line estimations.

Participants across conditions performed similarly when estimating events/objects one and four (“Life appears”/“Pluto”, “Humans appear”/“Cruithne”) ( $p > .05$ ); all participants were fairly accurate. For the temporal line estimations, participants who learned about both temporal and spatial magnitude hierarchically (THSH) were significantly more accurate when estimating event two “dinosaurs disappear” ( $t(74)=2.23, p=.03$ ) and event three “dinosaurs appear” ( $t(74)=2.6, p=.01$ ) on the time line than the conventionally only condition (TCSC).

Participants from the other conditions (THSC, TCSH, and TCSC) were not significantly different on this measure. Developing a more linear representation of temporal magnitude was associated with learning time and space hierarchically (THSH).

For the spatial line estimations, participants who learned spatial magnitude hierarchically were significantly more accurate when estimating object two “Mercury” (THSH:  $t(76)=3.04, p<.01$ ; TCSH:  $t(76)=3.16, p<.01$ ) and object three “Mars” (THSH:  $t(76)=3.47, p<.01$ ; TCSH:  $t(76)=3.74, p<.01$ ) on the number line than the conventionally only condition (TCSC). Participants who learned spatial magnitude hierarchically (THSH:  $t(76)=2.44, p=.02$ ; TCSH:  $t(76)=2.72, p=.01$ ) were also significantly more accurate than participants who learned temporal magnitude hierarchically and spatial magnitude conventionally (THSC) when estimating object three “Mars”; these conditions are not significantly different when estimating object two “Mercury” ( $p > .05$ ). Participants

who learned spatial magnitude conventionally (THSC and TCSC) are not significantly different when estimating objects two “Mars” or three “Mercury” ( $p > .05$ ). See Figure 4 for overall error on temporal and spatial estimations by condition. Developing a linear representation of spatial magnitude was associated with learning time or space hierarchically (THSH, THSC, or TCSH).

Table 2. Mean error (mm) by condition for content-specific number line estimations.

|      | Event one     | Event two           | Event three      | Event four   |
|------|---------------|---------------------|------------------|--------------|
|      | Humans Appear | Dinosaurs Disappear | Dinosaurs Appear | Life Appear  |
| THSH | 20.55(28.37)  | 43.09(31.61)        | 54.33(40.59)     | 20.81(20.63) |
| TCSH | 26.60(26.95)  | 69.45(41.12)        | 79.16(41.15)     | 25.32(30.05) |
| THSC | 33.17(41.23)  | 59.15(49.26)        | 73.53(50.87)     | 27.75(29.51) |
| TCSC | 31.66(28.53)  | 67.81(39.32)        | 86.35(40.78)     | 29.81(26.65) |
|      | Object one    | Object two          | Object three     | Object four  |
|      | Cruithne      | Mercury             | Mars             | Pluto        |
| THSH | 29.55(35.16)  | 28.32(40.65)        | 22.96(34.66)     | 18.11(26.65) |
| TCSH | 19.26(27.63)  | 23.43(34.65)        | 17.82(22.44)     | 29.32(38.04) |
| THSC | 22.12(25.06)  | 44.16(35.48)        | 48.03(41.60)     | 27.21(16.80) |
| TCSC | 30.90(24.66)  | 66.14(45.64)        | 61.76(35.99)     | 25.04(19.93) |

Note<sup>1</sup>: For each column, dark gray/white cells indicate a significant difference.

Note<sup>2</sup>: The light gray cells indicate no significant difference from any other group.

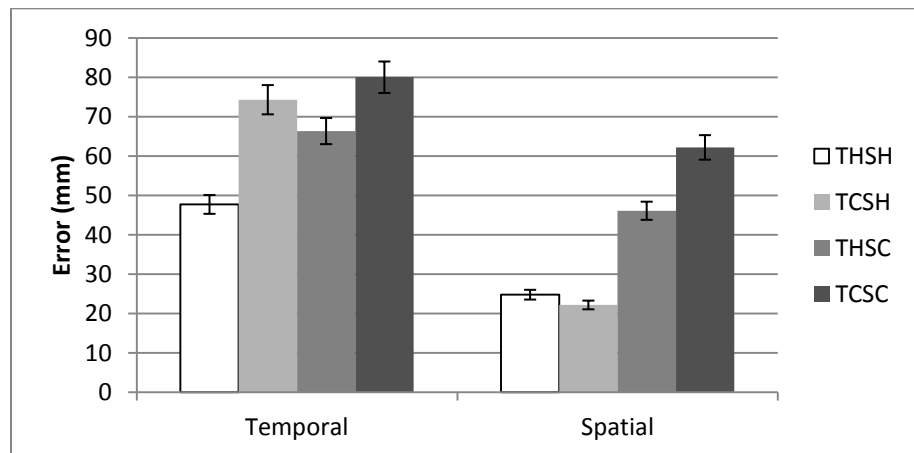


Figure 4. Average error (mm) across middle events/objects, respectively

*Development of scale assessing domain specific magnitude representation.*

Individual temporal and spatial magnitude line estimations were evaluated for the development of scale items to measure domain specific magnitude representation. Scale items were developed using items that had the strongest correlations and internal consistencies. For the temporal scale item, estimation of events two “dinosaurs disappeared” and three “dinosaurs appeared” were included in the scale item ( $r=.95$ ,  $p<.01$ ; Cronbach’s  $\alpha=.97$ ). For the spatial magnitude scale, estimation of objects two “Mercury” and three “Mars” were included in the scale item ( $r=.75$ ,  $p<.01$ ; Cronbach’s  $\alpha=.85$ ). The inclusion of event/objects two and three, and exclusion of event/objects one and four, are consistent with performance patterns described above: participants were all fairly accurate on the first and fourth items, with differences between conditions on estimation of items two and three. See Figure 4 for mean error on the temporal and spatial scale items by condition.

*Are temporal and spatial magnitudes represented in the same way?* Performance differences and similarities were identified on temporal and spatial magnitude estimation tasks, suggesting temporal and spatial magnitude representations have some common features. Analyses include examination of the success and effectiveness of the hierarchical alignment activity, the qualitative analysis of estimation patterns, and the quantitative analysis of error in estimation.

The hierarchical alignment activity was successful for developing linear representations of both temporal and spatial magnitudes. Participants who learned both temporal and spatial magnitudes hierarchically (THSH) were significantly more accurate on temporal and spatial magnitude estimations than participants who learned about temporal and spatial magnitudes conventionally (TCSC).

However, differences in effectiveness of the hierarchical alignment activity for temporal and spatial magnitudes suggest differences between temporal and spatial magnitude representation. Participants who learned only temporal magnitude hierarchically (THSC) were not significantly different from participants who learned temporal and spatial conventionally (TCSC); whereas participants who learned only spatial magnitude hierarchically (TCSH) were significantly more accurate than participants who learned about temporal and spatial magnitudes conventionally (TCSC).

Qualitative analysis of response patterns for temporal and spatial magnitudes suggests differences between temporal and spatial magnitudes. For temporal magnitude estimations, participants across conditions located the when “life appeared” accurately towards the bottom of the number line. Twenty-five (THSC) to thirty (THSH) percent of participants who learned about temporal magnitude hierarchically had correct representations of temporal magnitude (as defined by less than 20mm error for each estimation). Fifty percent of participants who learned about temporal and spatial magnitudes hierarchically (THSH) placed “humans appear” and “life appears” within 20mm of the correct response, and “dinosaurs appear” and “dinosaurs disappear” in the middle of wherever they located the other two events. The correct location for “dinosaurs appear” and “dinosaurs disappear” is within 3mm of the top of the time line (present day);

thus placing these items towards the middle of the time line is incorrect. The most common incorrect response (20%) for participants who learned about temporal magnitude hierarchically and spatial magnitude conventionally (THSC) compressed all events toward the bottom of the number line. For participants who learned about temporal magnitude conventionally (TCSH and TCSC), twenty percent of participants placed “humans appear” correctly at the top of the number line, and incorrectly placed the three remaining events (“life appear”, “dinosaurs appear”, and “dinosaurs disappear”) compressed at the bottom of the number line. No other patterns emerged from participants in these conditions (TCSH and TCSC). See Figure 5 for temporal magnitude response patterns.

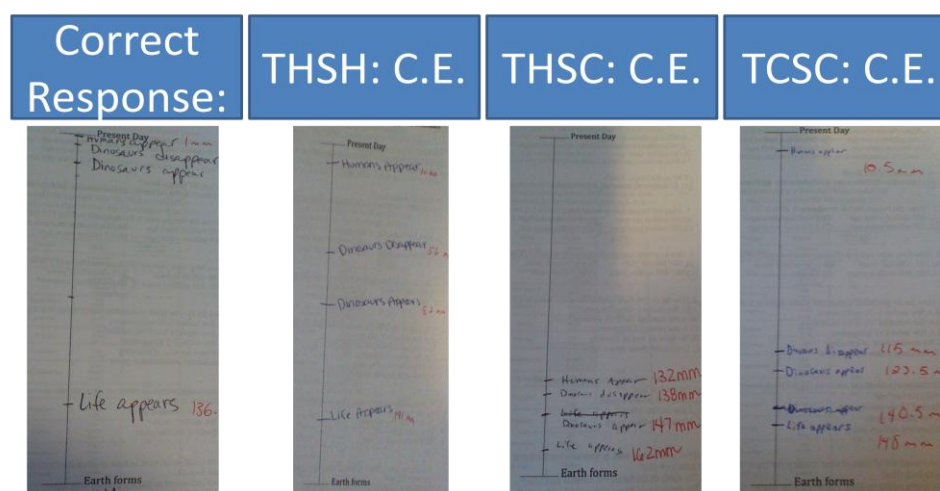


Figure 5. Response Patterns of Temporal Magnitude Estimation

Note: C.E. = Common Error, THSH = participants learn about temporal and spatial magnitudes hierarchically, TCSC = participants learn about temporal and spatial magnitudes conventionally, THSC = participants learn about temporal magnitudes hierarchically and spatial magnitudes conventionally

For spatial magnitude estimations, participants across conditions located “Pluto” fairly accurately towards the bottom of the number line, and group the remaining three objects (“Mars”, “Mercury”, and “Cruithne”) together. Estimation of the location of

“Mars”, “Mercury”, and “Cruithne” on the number line differed between conditions. Overall, 50% of responses from participants who learned a scale hierarchically (THSH, THSC, and TCSH) were correct (as defined by less than 20mm error for each estimation). Incorrect response options for participants who learned about spatial magnitude hierarchically (THSH and TCSH) tended to have “Mars”, “Mercury”, and “Cruithne” in a range of locations in the top third of the number line (~25%). The most common incorrect response for participants who learned about spatial magnitudes conventionally (THSC and TCSC) placed “Mars”, “Mercury”, and “Cruithne” in the middle third of the number line (~25%). See Figure 6 for spatial magnitude response patterns.

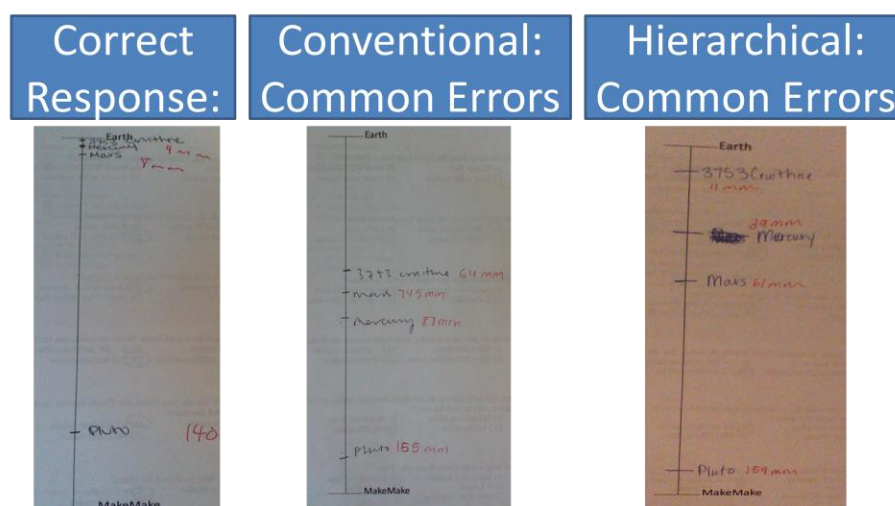


Figure 6. Response Patterns of Spatial Magnitude Estimations

Quantitative analysis of mean error for temporal and spatial magnitude estimations identifies both similarities and differences between temporal and spatial magnitudes. For both temporal and spatial estimations, participants were fairly accurate on estimation events and objects one and four; there were no significant differences when estimating object one “Cruithne” and event one “humans appear” ( $p > .05$ ;  $\mu$  difference=2.55,  $\sigma=38.01$ ), or object four “Pluto” and event four “life appeared” ( $p > .05$ ;

$\mu$  difference=2.14,  $\sigma$ =38.34). However, participants across conditions were more accurate when estimating object two “Mercury” than event two “dinosaurs disappear” ( $t(77)=3.55$ ,  $p<.01$ ;  $\mu$  difference=19.95,  $\sigma$ =49.64), and object three “Mars” than event three “dinosaurs appeared” ( $t(77)=6.16$ ,  $p<.01$ ;  $\mu$  difference=36.22,  $\sigma$ =51.97). Additionally, Participants across conditions were significantly more accurate on the spatially framed abstract (numeric) magnitude estimations than the temporally framed abstract magnitude estimations ( $t(77)=2.04$ ,  $p=.05$ ), although the actual difference between temporally and spatially framed abstract magnitude estimations was fairly small ( $\mu$  difference=4.11mm,  $\sigma$ =17.83mm).

Overall, differences in spatial and temporal estimation patterns in the effectiveness of hierarchical alignment activity, qualitative analysis of estimation patterns, and quantitative analysis of error for estimation of events/objects two and three may suggest differences in temporal and spatial magnitude representation. However, the overall structure of the hierarchical alignment activity is successful in developing a linear representation of both temporal and spatial magnitudes. Further, patterns of accuracy across individual estimations of events and objects follow similar patterns (i.e., increased accuracy for events/objects one and four, and increased variation for events/objects two and three). These similarities suggest that the nature of how people reason about different types of magnitude at different scales appears to be the same (i.e., the use of category boundaries to estimate magnitude). The difference in spatial and temporal estimations may be illuminating differences in domain knowledge; domain knowledge may influence number and placement of category boundaries, creating differences in representation.

*Generalization of magnitude: does learning a domain specific magnitude hierarchically transfer to abstract (numeric) magnitude representation?* There were eight number line estimations of abstract magnitude. Error on all eight line estimations were moderately to highly correlated ( $r$  ranging from .49 to .97,  $ps < .01$ ), and had high internal reliability (Cronbach's  $\alpha = .94$ ). Thus a scale was created using all abstract number line estimations (see Figure 7 for mean error).

There was borderline "high kurtosis" (5.8) in the condition where participants learned about temporal magnitude hierarchically and spatial magnitudes conventionally (THSC), and eight outliers across conditions: THSH=3 (<55mm in error), TCSH=2 (<90mm in error), and THSC=3 (<80 mm in error). Removing the outliers from analyses to examine if the data are sensitive to outliers reduces the kurtosis to a normal range (<3); however, overall findings are the same as when all participants are included in the analysis. Subsequently, the outliers were included in parametric analyses. Participants who learned at least one domain specific magnitude hierarchically were all significantly more accurate than participants from the TCSC condition (THSH:  $t(73) = 2.649$ ,  $p = .01$ ; THSC:  $t(73) = -2.084$ ,  $p = .04$ ; TCSH:  $t(73) = 2.086$ ,  $p = .04$ ). There were no significant differences between the conditions where participants learned at least one domain specific magnitude hierarchically (THSH, THSC, and TCSH) ( $p > .05$ ). This suggests that representation of domain-specific magnitudes transfers to abstract magnitude representation.

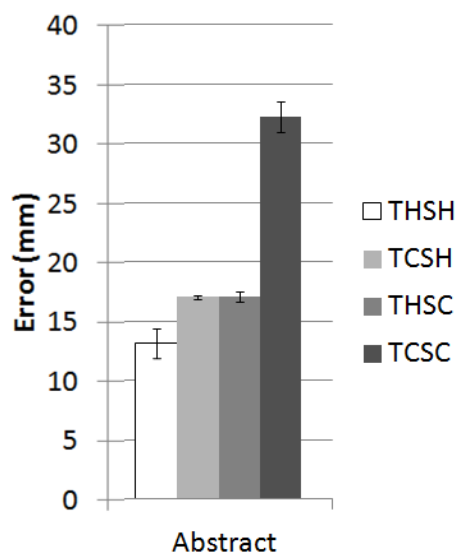


Figure 7. Overall mean error in abstract magnitude number line estimations by condition

*Are there differences between the representations of abstract magnitude at different scales?* Participants across conditions were significantly more accurate when estimating magnitudes on the million scale ( $\mu$  (error)=19.61,  $\sigma$ =26.40) than on the billion scale ( $\mu$  (error)=25.21,  $\sigma$ =31.80) ( $t(78)=2.61$ ,  $p=.01$ ). There are not enough magnitude estimations to statistically distinguish between models of magnitude representation (e.g., linear versus power functions with an exponent close to one). Siegler and Opfer (2003) suggest a minimum of 24 estimations, and the current study has eight. However, a visual analysis suggests estimations of magnitudes at both the million and billion scale follow the correct location linearly across conditions (see Figure 8). This finding is consistent with the segmented-linear model of magnitude representation (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008; Landy, et al., 2012). There are no significant differences between conditions on abstract magnitude estimations on the million scale ( $p>.05$ ). There is an effect of condition on the billion scale, with participants who learned at least one domain specific magnitude hierarchically making more accurate estimations than participants who learned temporal and spatial magnitudes conventionally (TCSC)

(THSH:  $t(77)=2.43$ ,  $p=.02$ ; TCSH:  $t(77)=2.12$ ,  $p=.04$ ; and THSC:  $t(77)=2.48$ ,  $p=.02$ ).

There were no significant differences between the THSH, TCSH, and THSC conditions ( $p>.05$ ). These findings suggest that while participants across conditions are able to accurately make estimations at the million scale, only those participants who learned temporal or spatial magnitude hierarchically tend to develop a linear representation of magnitude at the billion scale.

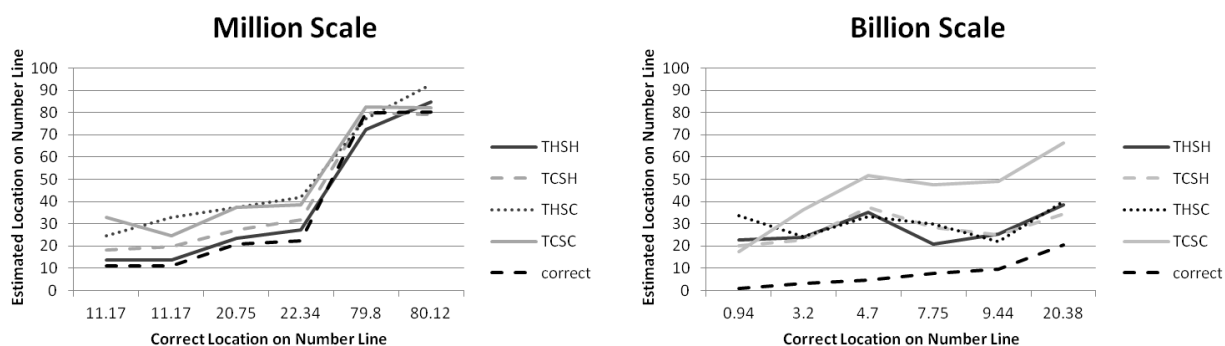


Figure 8. Estimated location on number line by correct location for magnitudes on the million and billion scales

*Does having linear representations of temporal or spatial magnitudes influence the understanding of scientific phenomena at related scales?* A series of multiple choice items were given as a measure of understanding of temporal and spatial content. There were no significant performance differences across conditions for the multiple choice items. Participants had significantly more items correct, on average, for the spatial content items than the temporal content items ( $t(71)=8.36$ ,  $p<.01$ ).

For each multiple choice question (e.g., “When did dinosaurs appear?”), there was a correct response (230 million years ago), and incorrect response options that ranged in difference in magnitude from the correct answer (e.g., 398 million years ago to 3.5 billion years ago). Thus, for each item, responses could be ranked from one (correct) to four

(incorrect by the largest amount). Performance on multiple choice items was not correlated with each other ( $p > .05$ ). Correlations were analyzed between having a linear representation of temporal and spatial magnitudes with the ranked responses to multiple choice items, using a Bonferroni correction for multiple comparisons. There were no significant correlations between performance on temporal and spatial magnitude estimations and performance on multiple choice items requiring recall only. This may suggest that domain specific magnitude representation is not used when answering multiple choice items requiring recall (alternative explanations are detailed in the discussion section). Performance on temporal magnitude estimations ( $r^2 < .425$ ,  $p^2 < .01$ ) were correlated with performance on four of the four multiple choice items that required recall plus reasoning. This suggests that temporal magnitude representation, at least in part, is used to understand temporal relationships between geologic events. Performance on spatial ( $r = .32$ ,  $p < .01$ ) magnitude estimations was correlated with performance on one of the four multiple choice items that required recall plus reasoning. On this item, participants are asked to identify how many miles Earth travels in a single year. That only one of four spatial items requiring recall plus reasoning was correlated with magnitude representation may be because participants were fairly accurate on spatial magnitude estimations; there is not enough variance in the spatial magnitude measure to detect correlations between spatial magnitude representation and understanding spatial relationships between objects.

*Do people recall categorical information versus numerical information*

*differentially?* Overall there was better recall of categorical facts about events and objects than corresponding numerical facts. For the eight multiple choice items assessing recall of categorical information versus numerical information, participants had significantly better recall for categorical information than numerical information for five of the eight items ( $t(78) < 8.00$ ,  $p < .01$ ). Participants had significantly better recall for numerical information than categorical information for one item ( $t(78) = 3.62$ ,  $p < .01$ ), and the remaining two items were not significantly different between numerical and categorical recall ( $p > .05$ ).

*Do prior mathematical skills, proportional reasoning skills, or magnitude*

*representations predict performance on large temporal and spatial magnitude*

*representation?* Participants were assigned a ranking of “poor at math”, “average at math”, or “strong at math”. Eleven participants were labeled as “poor at math”, thirty-seven as “average at math”, and twelve as “strong at math”. The relatively low number and unequal sample sizes preclude statistical tests of significance; however, as a preliminary analysis, differences in average error between mathematical skill levels were assessed. Participants labeled as “poor at math” had the greatest amount of error when estimating temporal ( $\mu = 103.38$ ,  $\sigma = 29.58$ ), spatial ( $\mu = 34.58$ ,  $\sigma = 28.01$ ), and abstract ( $\mu = 23.57$ ,  $\sigma = 21.90$ ) magnitude, participants labeled as “average at math” had less amount of error (temporal:  $\mu = 59.19$ ,  $\sigma = 40.64$ , spatial:  $\mu = 30.22$ ,  $\sigma = 34.49$ , abstract:  $\mu = 15.04$ ,  $\sigma = 17.25$ ), and those labeled as “strong at math” had the least amount of error (temporal:  $\mu = 33.88$ ,  $\sigma = 30.82$ , spatial:  $\mu = 21.85$ ,  $\sigma = 36.60$ , abstract:  $\mu = 3.82$ ,  $\sigma = 1.33$ ). This may suggest that having a linear representation of magnitude is associated with mathematical

skills, or that the hierarchical alignment activity, a mathematically based intervention, is not as effective for people with lower mathematical skills.

All participants possessed basic proportional reasoning skills, with all participants answering at least four out of five proportional reasoning questions correctly. No demographic information (e.g., sex, age, handedness, etc) was related with performance ( $p > .05$ ), including previous participation in geoscience courses ( $p > .05$ ).

Participants completed an abstract (numeric) line estimation task prior to participating in a condition. Pretest abstract magnitude estimations were all moderately correlated (mean  $r = .69$ ,  $p < .01$ ), and had strong internal reliability (Cronbach's  $\alpha = .93$ ). Thus, all pretest abstract magnitude estimations were included in a scale item. Pretest abstract magnitude estimation was not correlated with estimations of temporal or spatial magnitude across conditions ( $p > .05$ ). This suggests a distinction between abstract magnitude representation and domain-specific magnitude representation.

## Discussion

The hierarchical alignment activity was successful in developing a linear representation of domain specific magnitude. Participants provided with hierarchically structured magnitude information were more accurate on number line estimations than participants given the same content in a conventional manner for temporal and spatial magnitudes. These findings are aligned with the Category Adjustment Model (CAM), suggesting people use hierarchically organized categorical information when making estimations across scales and across dimensions; and that providing people with more salient category boundary information improves estimation.

Participants across conditions performed similarly when estimating event/object one and four (“Humans appear”/“Cruithne”, “Life appears”/“Pluto”). This may be because these events/objects are closest to the number line flanks, and thus are anchored by the relatively close flanks of the number line itself (‘top’ and ‘bottom’). Whereas event/object two and three (“Dinosaurs disappear”/“Mercury”, “Dinosaurs appear”/“Mars”) may not be naturally perceived in these same salient categories; they are located ‘somewhere in between’. Consistent with this interpretation, participants from the conventional only condition demonstrate more bias in estimation towards the center of the number line than the participants from the hierarchical conditions. This finding is aligned with the three-category representation of geologic time advocated by Trend (2001), as well as predictions of biases towards the middle of these categories by the CAM (Huttenlocher, et al., 1988). However, more research is needed to further identify and characterize categories used in the representation of geologic time and astronomical distances.

Performance on temporal, spatial, and abstract (numeric) magnitude measures had similarities and differences. Participants across conditions were the most accurate when estimating abstract magnitudes, then spatial magnitudes, and least accurate on temporal magnitudes. There were qualitative differences between temporal and spatial magnitude representations, with events erroneously compressed towards the bottom of the number line and objects erroneously compressed at the center of the number line.

While the hierarchical alignment activity was successful in developing a linear representation for both temporal and spatial magnitudes, there were also differences in the effectiveness in developing a linear representation between temporal and spatial

magnitudes. Only participants who were presented with both temporal and spatial magnitudes hierarchically (THSH) were significantly more accurate on temporal magnitude estimations compared with the conventionally only condition; whereas, only spatial magnitudes needed to be presented hierarchically (THSH and TCSH) for participants to be significantly more accurate on spatial magnitude estimations. Further, there is some evidence of transfer from learning temporal magnitudes hierarchically to spatial magnitude representation when spatial magnitude is presented conventionally (THSC): seven participants in the THSC condition had the correct representation for spatial magnitude.

One explanation for this pattern of differences in performance is that temporal, spatial, and abstract magnitudes are represented differently (see Agrillo, et al., 2010 and Walsh, 2003 for a discussion on the existence of a general magnitude system). Alternatively, it may be the case that temporal, spatial, and abstract magnitudes are all represented in a similar way, but preexisting knowledge (and misconceptions) bias the subjective categories people use to make estimations. For example, consistent with participants being better at estimating spatial magnitudes compared to temporal magnitudes, that geologic time is often neglected in the classroom (Dodick, 2007; Trend, 2001) and learning about the solar system is commonplace, it seems likely participants did have more knowledge of the solar system than geologic time. Future research should examine unfamiliar scales, both in content and magnitude. For example, one may use an unfamiliar solar system, which would have a different time-course as well as different celestial objects.

Differences in effectiveness of the hierarchical alignment activity for spatial and temporal magnitudes may also be due to familiarity: the first three base analogies (tens, hundreds, thousands) may be differentially familiar to participants for temporal and spatial magnitudes. While temporal and spatial scales of magnitude were aligned, participants may be more familiar with traveling tens, hundred, and even thousands of miles; whereas participants could have only personally experienced years at the tens scale (no participants were over one hundred years old). Increased familiarity with multiple base concepts may have made the overarching spatial magnitude analogy more salient compared with the overarching temporal analogy (which is predicted by the principles of progressive alignment (Kotovsky & Gentner, 1996)).

It is worth briefly discussing the finding that participants who learned temporal magnitude hierarchically and spatial magnitude conventionally (THSC) did not develop a linear representation of temporal magnitude. One possible explanation is that learning a second magnitude (e.g., space) conventionally interferes with recall of the first magnitude (e.g., time). Alternatively, two examples may be required in this context to develop a linear representation of temporal magnitude. Learning with two examples has been found to be more effective than learning with just one example (Gentner, Loewenstein, & Thompson, 2003). Studies presenting content specific magnitudes in different orders, as well as different numbers of examples, are required to explore these two possible explanations.

Findings from the abstract numerical magnitude task are consistent with the segmented number line model of scale representation (Ebersbach, et al., 2008; Landy, et al., 2012). The segmented linear model posits separate linear functions for sets of magnitudes (e.g., ‘familiar’ versus ‘unfamiliar’) when estimated magnitude is plotted against actual magnitude. For example, Ebersbach, et al. (2008) found young children had a fairly accurate linear slope for smaller, familiar numbers, and a separate shallower linear slope for larger, unfamiliar numbers. While there were not enough estimations in the current study to carefully characterize the slope function, findings show estimation patterns consistent with the segmented linear model: participants across conditions had a more accurate linear slope for estimations made on the million scale, and, while still linear, were significantly less accurate on estimation on the billion scale (overestimation). More research is needed examining estimations at large scales for detailed modeling of these slope functions.

That the hierarchical condition transferred to estimations about abstract numerical magnitudes, suggests that people use categorical information when making these types of estimations. While there are some studies that look at the subjective categorization of numbers (Laski & Siegler, 2007; Mix, Huttenlocher, & Levine, 2002; Siegler & Robinson, 1982), there has not been previous work mapping the CAM onto number line estimations and scale representation. While more direct and explicit research is needed, we speculate that the CAM could serve as a unifying model for currently competing theories (e.g., logarithmic-to-linear, power function with anchor points, segmented linear). Category boundaries may serve as distinct anchor points, with adults possessing more precise categories (at the individual numbers level) compared with children.

Whereas young children may have many numbers in one “big” or “unfamiliar” category, adults may possess counting strategies for numbers within “unfamiliar” scales. Thus, the CAM offers an account for the overestimation of unfamiliar magnitudes that maintain linearity within the scale. More extensive research is needed to identify types of categories used in scale representation to see if a CAM can predict the changing pattern of bias in number line estimations that occurs with development.

There were no significant differences between conditions on any multiple choice item. While it is possible these items do not assess understanding of scale information; this explanation seems unlikely because the items explicitly require magnitude information to answer correctly. Another possible explanation is the study was underpowered to address these particular questions. Given the 20 participants in each condition, and using the standard 80% power, an effect size (measure by Cohen’s  $d$ ) of .8 to 1 is required to detect performance differences between conditions. This range of effect sizes is fairly large (for medium effect size, Cohen’s  $d = .5$ ). The small effects sizes found in this study (Cohen’s  $d < .1$ ) would not be detectable.

Another explanation for why the intervention and control conditions did not differ on recall-based multiple choice items is because there is sufficient learning in both conditions to answer these particular items. The control condition presented information about the scientific phenomenon using progressive alignment, structural alignment, and multiple opportunities to make analogies. Thus, it is reasonable to suspect enough learning took place to answer multiple choice questions. However, most of the multiple choice questions were not at ceiling. Of the 14 temporal questions, three items had above 86% accuracy, four items had below 24% accuracy, with the remaining seven items

having more variance. Of the 14 spatial questions, six items had between 71% and 85% accuracy, with the remaining items having more variance. This suggests that only for particular items there was learning across conditions, and for other items there was not. Accuracy on some multiple choice items and not others may be driven by salient pieces of information learned during the activities. Anecdotally, participants seemed interested to learn dinosaurs appeared in the Triassic, because, as noted by the participants, participants knew of the movie “Jurassic Park”. Related, 94% of the participants recalled correctly that dinosaurs appeared in the Triassic.

A final possibility for there being no differences between conditions on the multiple choice questions is, while the hierarchical alignment activity may have developed more linear representations of magnitude, novices may not use magnitude information to estimate their responses. The participants may have approached these questions as all-or-nothing; they either knew the answer or they did not. Interestingly, the nature of geoscience is to use current day spatial configurations of strata (rock layers) to estimate a sequence of events (Parcell & Parcell, 2009). Thus, this may represent expert/novice differences in reasoning about magnitude information, where experts are inclined to use knowledge to make estimations of unknown magnitudes whereas novices are not. However, the format of the question (i.e., having discrete response options) may have, at least in part, promoted an all-or-nothing approach. An open-ended format, such as number line estimations, may help promote estimation-based responses by novices, allowing for comparison of magnitude representation in estimation of scientific phenomenon. Number line estimations may also be more sensitive measures to variance in student performance to combat issues of low power.

There were two types of multiple choice questions: questions that required magnitude recall only and those that required magnitude recall plus an additional step of reasoning. For both temporal and spatial questions, temporal, spatial, and abstract magnitude representation is not associated with accuracy on items that required magnitude recall only. Representation of temporal magnitude was associated with more accurate responses on all temporal items that required magnitude recall plus reasoning. Representation of spatial magnitude was associated with more accurate responses on one of the spatial items that required magnitude recall plus reasoning. This item asked participants how many miles Earth moves in a single year. A possible explanation why all the spatial multiple choice questions requiring magnitude recall plus reasoning were not related to spatial representation is because participants were fairly accurate on spatial magnitude estimations overall. Thus, this would suggest that more sensitive measures are required to detect differences and correlations for spatial estimations. That the temporal questions requiring magnitude recall plus reasoning were related to temporal magnitude representation suggests that participants, at least in part, use temporal magnitude to estimate their answer.

For each of the recall only multiple choice questions, there were two types of response options: half of the questions had categorical response options, and half had numerical response options. Overall there was better recall of categorical facts about events and objects than corresponding numerical facts. This suggests that for temporal and spatial magnitudes outside of human perception, categories of magnitude may be more salient than the actual magnitudes. There are a number of reasons why categories may be more salient than numerals for large temporal and spatial magnitudes. If

magnitudes are represented using category boundaries (e.g., Crawford, et al., 2006; Huttenlocher, et al., 1988; Newcombe, et al., 1999), and participants had few categories for large unfamiliar magnitudes (Siegler & Opfer, 2003; Trend, 2001), it may be harder to discriminate between individual numerals within one conceptual category (“big”) than to discriminate between discrete words. Alternatively, differences in character length may have influenced the differences in recall of categories versus numbers (e.g., “Cenozoic” category is eight characters whereas the corresponding numeral response option “65 million years ago” is twenty characters). Shorter words are recalled with more accuracy than longer words (Cowan, Baddeley, Elliott, & Norris, 2003). Positioning of category and numeral labels in the hierarchical and convention activities may also lead to biased recall. For the hierarchical condition, each number line was labeled with categorical information above the number line, whereas numerals were labeled on the number line itself. This type of presentation may highlight category labels over numeral labels. For the conventional condition, categories were always listed on the left side and number listed on the right, which may facilitate a primacy effect (greater recall for items presented first) (Ebbinghaus, 1913). More research is needed to characterize recall of categories versus numeral for temporal and spatial scales outside of human perception. Additionally, future research may examine the category versus numeral recall across scales.

The hierarchical alignment activity is based on a number of principles: progressive alignment, hierarchical alignment, structural alignment, and multiple opportunities to make the analogy. The current study measured the additive benefit of one of these principles: hierarchical alignment. The intervention condition presented scale

information hierarchically and progressively, whereas the conventional condition presented scale information progressively. That participants who learned scale information hierarchically are more accurate than participants who learned scale information conventionally, suggests there is an added benefit of the hierarchical alignment of scale information. The role of having scale information presented progressively was beyond the scope of this dissertation. Future research should systematically alter the principles used in the hierarchical alignment activity to identify individual contribution. Such findings are important for the theoretical development of analogical reasoning, as well as successful implementation as learning tools in classrooms.

The current study finds an additive benefit of the using hierarchical alignment to teach magnitude compared with an activity that only uses principles of progressive alignment. This suggests that having the opportunity to engage in hierarchically aligning magnitudes at different scales is important for developing a linear representation of magnitude, and, thus, has programmatic implications for curriculum design. Analogy and visual displays are the most commonly used pedagogical practices when teaching about magnitudes outside of human perception (Libarkin, et al., 2007). However, there are a number of potential barriers to alignment, such as unfamiliar base concepts, dissimilar base and target concepts, psychological barriers, and practical constraints of the classroom (see chapter one for detailed discussion). The current study illustrates the benefit of hierarchical alignment in addition to progressive alignment in learning magnitude information through analogy. Additionally, the current study finds transfer from learning domain-specific magnitudes hierarchical to developing a more linear

abstract (numeric) magnitude representation. As having a linear representation of scale is predictive of performance on a range of standardized tests in mathematics (Siegler & Booth, 2004), transfer between domain-specific and abstract magnitudes suggests hierarchical alignment may also be useful in learning about magnitudes outside of human perception in mathematical classes.

## CHAPTER 3

### CORRECTIVE FEEDBACK

#### Aims

The prior study demonstrated that the hierarchical alignment activity is an effective instructional tool for teaching large temporal, spatial, and abstract (numeric) magnitudes. A drawback of the hierarchical alignment activity is that, while effective, the activity may not be efficient enough to be utilized in an already packed curriculum (it takes 1.5 hours to complete). Therefore, an aim of the current study is to develop an alternative activity that can be more naturally integrated into normal lectures in geoscience courses (the corrective feedback activity), by scientifically assessing the principles of the hierarchical alignment activity.

The hierarchical alignment activity is based on a number of principles: giving multiple opportunities to align time to space in a linear representation, using the same amount of space for each alignment (structural alignment), progressing from small familiar scales to geological scales, and hierarchically organizing all previous scales within the current scale. It is not clear if all principles are required for a successful analogy, or, if, some principles are more important than others. For example, the prior study (Transfer study) found that including progressive and hierarchical alignment of scale information is more effective than progressive alignment alone. The current study investigates the necessity of two of the principles of the hierarchical alignment model: the progressive and hierarchical alignment of scale information.

An activity referred to as the ‘corrective feedback activity’ was developed based on a select subset of principles of the hierarchical alignment activity: the corrective

feedback activity gives students multiple opportunities to align time and space in a linear representation and uses the same amount of space for each alignment. Importantly, the corrective feedback activity differs from the hierarchical alignment activity by not progressing from small familiar scales to geological scales and not hierarchically organizing all previous scales within the current scale. Rather, the corrective feedback activity provides students with corrective feedback on magnitude estimations. Corrective feedback has been found to be effective for learning about unfamiliar magnitudes (Thompson & Opfer, 2010).

Directly comparing the hierarchical alignment activity and the corrective feedback activity, and the individual principles on which they are based, is beyond the scope of this dissertation. However, if the corrective feedback activity is successful in developing a linear representation of geologic time, it would suggest that progressing from small familiar scales to geologic scales and hierarchically organizing all previous scales within the current scale are not necessary components of developing a linear representation of geologic time. If the corrective feedback activity fails to develop a linear representation of geologic time, it would suggest that either one or both of these principles are necessary for developing a linear representation of geologic time. Further research would be needed to examine the individual contribution of each principle.

The current study included a course that was taught by two instructors as separate classes. While one instructor administered the corrective feedback activity as prescribed, the other instructor did not. The instructor who did not administer the corrective feedback activity as prescribed (i.e., provide corrective feedback), provided the corrective feedback information as a linear representation of the Geologic Time Scale. Thus, I will refer to one class as the “corrective feedback” class and the other as the “linear visualization” class. See the Procedure section below for a detailed description of corrective feedback and linear visualization.

## Methods

### Participants

Participants consisted of all students enrolled in an undergraduate introductory-level geoscience course aimed primarily at non-majors during the Fall 2011 and Spring 2012 semesters at a large urban American university. In Fall semester 2011, there were 98 students enrolled in the corrective feedback class and 108 enrolled in the linear visualization class. In Spring semester 2012, there were 96 students enrolled in the corrective feedback class and 94 enrolled in the linear visualization class. See Table 3 for demographics of total enrollment.

Table 3. Demographics of total enrollment by class and condition

|                                  | Frequency (%)                |                           |                               |                           |
|----------------------------------|------------------------------|---------------------------|-------------------------------|---------------------------|
|                                  | Corrective Feedback          |                           | Linear Visualization          |                           |
|                                  | Control                      | Intervention              | Control                       | Intervention              |
| Age                              | $\mu=22.1$ ,<br>$\sigma=1.4$ | $\mu=21.7$ , $\sigma=1.9$ | $\mu=22.72$ ,<br>$\sigma=1.9$ | $\mu=21.7$ , $\sigma=1.3$ |
| Sex                              |                              |                           |                               |                           |
| Male                             | 45 (43.3)                    | 71 (59.7)                 | 55 (45.8)                     | 44 (41.5)                 |
| Female                           | 59 (56.7)                    | 48 (40.3)                 | 65 (54.2)                     | 62 (58.5)                 |
| Race                             |                              |                           |                               |                           |
| White/Caucasian                  | 81 (77.9)                    | 89 (74.8)                 | 82 (68.3)                     | 67 (63.3)                 |
| African-American                 | 10 (9.6)                     | 8 (6.7)                   | 13 (10.8)                     | 12 (11.3)                 |
| Asian                            | 4 (3.8)                      | 8 (6.7)                   | 7 (5.8)                       | 6 (5.7)                   |
| Hispanic                         | 4 (3.8)                      | 5 (4.2)                   | 5 (4.2)                       | 10 (9.4)                  |
| Not Identified                   | 4 (3.8)                      | 8 (6.7)                   | 7 (5.8)                       | 8 (7.5)                   |
| Other                            | 1 (1.0)                      | 1 (0.8)                   | 6 (5.0)                       | 3 (2.8)                   |
| Education Level                  |                              |                           |                               |                           |
| Freshman                         | 16 (15.4)                    | 21 (17.6)                 | 29 (24.2)                     | 9 (8.5)                   |
| Sophomore                        | 54 (51.9)                    | 49 (41.2)                 | 42 (35.0)                     | 48 (45.3)                 |
| Junior                           | 27 (26.0)                    | 33 (27.7)                 | 35 (29.2)                     | 29 (27.4)                 |
| Senior                           | 7 (6.7)                      | 9 (7.6)                   | 12 (10.0)                     | 14 (13.2)                 |
| 5+                               | 0 (0.0)                      | 0 (0.0)                   | 2 (1.7)                       | 6 (5.7)                   |
| Degree                           |                              |                           |                               |                           |
| Bachelor of Arts                 | 70 (67.3)                    | 62 (52.1)                 | 56 (46.7)                     | 58 (54.7)                 |
| Bachelor of Science<br>Education | 6 (5.8)                      | 12 (10.1)                 | 13 (10.8)                     | 11 (10.4)                 |
| Bachelor of Business             | 5 (4.8)                      | 23 (19.3)                 | 21 (17.5)                     | 13 (12.3)                 |
| Bachelor of Science              | 16 (15.4)                    | 18 (15.1)                 | 20 (16.7)                     | 14 (13.2)                 |
| Other Bachelor<br>degree         | 7 (6.8)                      | 4 (2.5)                   | 10 (8.3)                      | 9 (7.5)                   |
| Non-degree                       | 0 (0.0)                      | 1 (0.8)                   | 0 (0.0)                       | 1 (0.9)                   |

Note: not all percentages to sum to 100% as some student data was unavailable

### Procedure

In Fall 2012, students received regular instruction (control group). Regular instruction was administered primarily through lecture via PowerPoint presentation. Geologic events were taught over approximately ten class sessions. Lectures on geologic events were organized by temporal divisions in the Geologic Time Scale (e.g., one lecture was on the Hadean Eon, another on the Archean Eon). At the beginning of each new

lecture, the image of the Geologic Time Scale from the class text book was presented as an introduction, with the relevant division in time highlighted (see Figure 9). Importantly, the image of the Geologic Time Scale is not linear; the Precambrian (roughly 4 billion years) is compressed to the bottom of the time line, and the Phanerozoic Eon (542 million years) is expanded. This representation of the Geologic Time Scale is conventional and functional as it allows users to see the smaller divisions of time in the Phanerozoic. Both teachers estimated presentation of this introductory slide was approximately ten seconds.

After learning about geologic events over the ten class sessions, students were tested on this material on a noncumulative exam (exam three) in the next class session. To assess magnitude representation and reasoning eleven multiple choice items and four line estimation items were added to the regular exam (see Measures section below for description of items). While there may have been differences between lectures in the corrective feedback and linear visualization classes, the instructors worked together toward developing similar classes. The instructors shared course materials, lecture slides, and exam questions.

**Phanerozoic Eon  
Paleozoic Era  
Cambrian & Ordovician Periods  
(Early Paleozoic)**

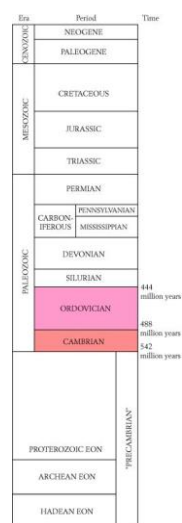


Figure 9. Introductory slide presented in normal class instruction

Note<sup>1</sup>: Both the corrective feedback class and the linear visualization class saw this slide.

In the Spring 2012 semester the corrective feedback activity was administered (intervention group). In the corrective feedback activity, the introductory slide with the Geologic Time Scale image was replaced with two different PowerPoint slides. The first slide was comprised of the same Geologic Time Scale image; however, there was also a blank time line to the right of the original image (see Figure 10). Four locations on the blank time line were labeled A through D. Students were asked which of the four response options showed where the highlighted division in time would be located on the linear scale. Once students responded, the instructor presented the second slide, giving the students corrective feedback (see Figure 11). On the corrective feedback slide, the original image was presented, along with an image of a linear representation of the Geologic Time Scale. The relevant divisions in time were highlighted on both images and arrows connected the highlighted divisions in time on both images.

The corrective feedback activity was administered differently in the corrective feedback and linear visualization classes. The clicker response system (students responded to questions via a hand-held electronic device) was used for the corrective feedback activity. On the first slide where students responded to the question of where is the relevant division of time on the linear scale, students from the corrective feedback class responded via the clicker response system. The linear visualization teacher does not use the clicker response system. Thus, students from the linear visualization class were supposed to respond by raising their hand when the correct response option was given. Unfortunately, due to a communication error, slide one was not shown at all. Rather, the linear visualization class was only presented with the corrective feedback slide as a linear visualization of the Geologic Time Scale. Corrective feedback teacher estimated the

presentation of both slides took on average thirty seconds, with the first few presentations requiring more time because students were given initial instructions and explanation of what the image represented. Linear visualization teacher estimated the presentation of the one slide took on average ten seconds, with the first few presentations requiring more time because students were given initial explanation of what the image represented.

Today we are going to talk about the Cambrian & Ordovician Periods.

You can see them highlighted on the Geologic Time Scale.

Where would these periods be located on the linear time scale on the right?

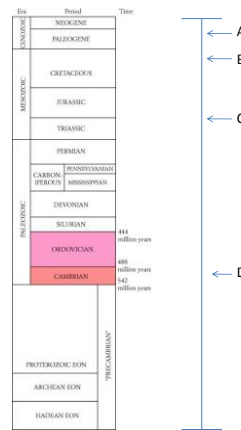


Figure 10. Example of a clicker slide.

Note<sup>1</sup>: Students are asked to locate where the highlighted divisions on the Geologic Time Scale would be located on the linear time line.

Note<sup>2</sup>: only the corrective feedback class saw this slide.

Phanerozoic Eon  
Paleozoic Era  
Cambrian & Ordovician Periods  
(Early Paleozoic)

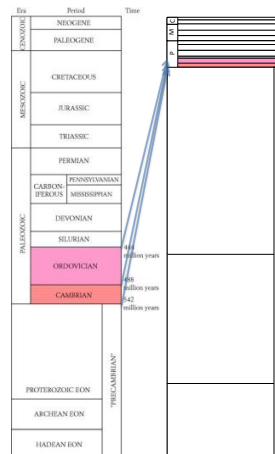


Figure 11. Example of slide containing corrective feedback.

Note<sup>1</sup>: The conventional Geologic Time Scale is aligned with the linear time scale.

Note<sup>2</sup>: Both the corrective feedback class and the linear visualization class saw this slide.

## Measures

To measure understanding and representation of temporal magnitude, students completed the same 14 multiple choice items as in the Transfer study (see Appendix A). However, not all students were given all 14 items. Eight of these 14 items require magnitude recall only (e.g., “When did dinosaurs appear?”). To measure differences in recall of categorical and magnitude information, four items included category response options (e.g., “A. Triassic...”), and four included analog abstract (numeric) magnitude response options (e.g., “A. 230 million years ago...”) (230 million years ago was the Triassic). Two exams versions were created; with version one having two items with categorical response options and two different items with magnitude response options, and exam version two having the opposite set of items. This between-subjects design was adopted because of limited space on the exam for additional items and to maintain a normal exam experience (i.e., repeated items are atypical on a classroom exam). Both exam versions had the other seven temporal multiple choice items from the Transfer study (see Appendix A). Four items required magnitude recall plus an additional step of comparison (e.g., “What is the relationship between dinosaurs disappearing and humans appearing?”). These items were developed through collaboration between a cognitive psychologist and a geologists specializing in scientific phenomenon that occur at large scales. There were two items were from the Geoscience Concept Inventory (Libarkin, et al., 2005). In one item, students were asked to choose the correct linear time line from five response options. In the other item, students were asked to choose the statement that best described the Earth when it first formed. A final item was developed by Barghaus and Porter (2010) for use with middle school students. In this item, students use an image

of the Geologic Time Scale to identify the true statement. Despite being provided with precise magnitudes, the true statement (“The Proterozoic lasted much longer than the Phanerozoic”) is counterintuitive because the amount of space provided on the Geologic Time Scale is not aligned with the magnitudes. Thus, each student answered eleven multiple choice items on temporal magnitude in addition to regular exam questions. The time line item from the GCI was designed to measure geologic time representation. The other multiple choice items were designed to measure knowledge and reasoning about geologic time.

To measure abstract (numeric) magnitude representation, students completed the same abstract magnitude number line estimations as the Transfer study (see Appendix A).

There were three non-cumulative exams given during the semester; exam grades were collected. Exams one and two were taken prior to the corrective feedback activity being introduced; thus, exams one and two serve two functions. They measure how similar the students are from the Fall 2012 semester (control) and Spring 2013 semester (intervention), and they also measure how similar the corrective feedback class and the linear visualization class are unrelated to the intervention. Exam three grades measure effectiveness of the corrective feedback activity. Absences for the semester were also collected. Demographic information was obtained through the General Education Office.

## Results

In order to account for exposure to intervention, students who attended at least seven of the ten classes (70% attendance rate) while the intervention was being administered were included in analysis (and equivalent classes in the control condition). One hundred and forty-nine out of 194 students (75 in control group) were included in

analysis from the corrective feedback class, and 111 of 202 (61 in control group) were included from the linear visualization class. While roughly 50% of the linear visualization class was excluded from analysis, 71% of the excluded students were absent for more than half of the classes during the intervention.

*Are classes A and B similar prior to intervention?* There are two pretest measures (exam one and two). The corrective feedback class had significantly lower grades than the linear visualization class on exam one ( $t(258)=4.03, p<.01$ ). The two classes were not significantly different on exam two. Given the inconsistent differences between the corrective feedback and linear visualization classes, as well as the difference in intervention administration, analyses will examine each class separately.

*Are the control and intervention groups similar prior to intervention?* For the corrective feedback class, the intervention group had significantly higher grades on exams one ( $t(147)=3.17, p<.01$ ) and exam two ( $t(147)=2.21, p=.03$ ) than the control group. This suggests that for the corrective feedback class, the intervention group is performing better in this course overall compared with the control group unrelated to the intervention. For the linear visualization class, there are no significant differences between the intervention and control groups on exams one and two ( $p>.05$ ). This suggests that for the linear visualization class, the intervention and control group are not significantly different from one another.

*Development of abstract (numeric) magnitude scale.* Two scale items were developed for assessing abstract magnitude representation at the billion and million scales. There were eight number line estimations of abstract magnitude, with four estimations at the billion scale and four estimations at the million scale. Error on all four

abstract line estimations at the billion scale were highly correlated with each other (mean  $r = .85$ ,  $ps < .01$ ). Three of the four abstract line estimations at the million scale (35 million, 65 million, and 70 million) were moderately correlated with each other (mean  $r = .74$ ,  $ps < .01$ ). The remaining estimation at the million scale (251 million) was less correlated with the other estimations at the million scale (mean  $r = .45$ ,  $ps < .01$ ). Estimations at the billion scale are moderately correlated with the first three estimations at the million scale (mean  $r = .51$ ,  $ps < .01$ ), and weakly correlated with the estimation of 251 million (mean  $r = .26$ ,  $ps < .01$ ). Participants are significantly more accurate when estimating 251 million out of 542 million than all other line estimations ( $\mu$  difference = 11.17,  $\sigma = 28.62$ ,  $ps < .01$ ). However, removing the estimation of 251 million out of 542 million from analyses (as part of a sensitivity analysis) does not change outcomes. Thus, because the estimation of 251 million is correlated with the other estimations (albeit weakly), and exclusion does not alter outcomes, further analyses will include all line estimations.

There was strong internal validity across all line estimations (Cronbach's  $\alpha = .88$ ); however, when estimations from billion and million scales are isolated there is a divergence in internal validity: estimations at the billion scale increase to Cronbach's  $\alpha = .92$  and estimations at the million scale decrease to Cronbach's  $\alpha = .78$ . Cronbach's  $\alpha$  can be inflated with an increase of items (even with low inter-item reliability, the more items in the overall scale, the higher the Cronbach's  $\alpha$ ). Thus, that the reliability of the billion scale increases when separated from the million scale is particularly suggestive that there are meaningful differences between the representation of magnitudes at the billion and million scales.

Consistent with this interpretation, as described above, there are also differences in correlations between the billion and million scales. Thus, the current study will adopt two scale items for assessing abstract magnitude representation at the billion and million scales.

*Is the intervention successful? The corrective feedback class:* After statistically controlling for their superior performance on exams 1 and 2 (using an ANCOVA), students in the intervention group performed better on exam 3 than the control group ( $f=6.44$ ,  $p<.01$ ,  $\text{cohen's } d=.5881$ ) (see Figure 12 for means). Participants from the intervention condition were more accurate on the time line item from the GCI (which measures temporal magnitude representation) than the control condition ( $\chi^2(1)=3.64$ ,  $p=.04$ ). This suggests that participants from the intervention condition had a more linear representation of geologic time than participants from the control condition. There were no significant differences on the other multiple choice items that measure knowledge and reasoning about temporal magnitude ( $p>.05$ ). Overall accuracy on the multiple choice items ranged from 77% to 44%. Students in the intervention group were significantly more accurate when estimating abstract magnitudes at the billion scale ( $t(110)=3.43$ ,  $p<.01$ ). There were no significant differences between the intervention and control group when estimating abstract magnitudes at the million scale ( $p>.05$ ). See Figure 13 for means of abstract magnitude line estimations. Abstract magnitude representation does not predict performance on exams ( $p>.05$ ). There are no significant differences between categorical and magnitude recall ( $p>.05$ ).

**The linear visualization class:** Students in the intervention group performed significantly worse on exam 3 than control group ( $t(109)=2.98$ ,  $p<.01$ ). There were no

significant differences on any of the multiple choice items ( $p > .05$ ). Overall accuracy on the multiple choice items ranged from 76% to 82%. There were no significant differences between the intervention and control group when estimating abstract magnitudes at either the billion or million scales ( $p > .05$ ). Abstract magnitude representation did not predict performance on exams ( $p > .05$ ). There were no significant differences between categorical and magnitude recall ( $p > .05$ ).

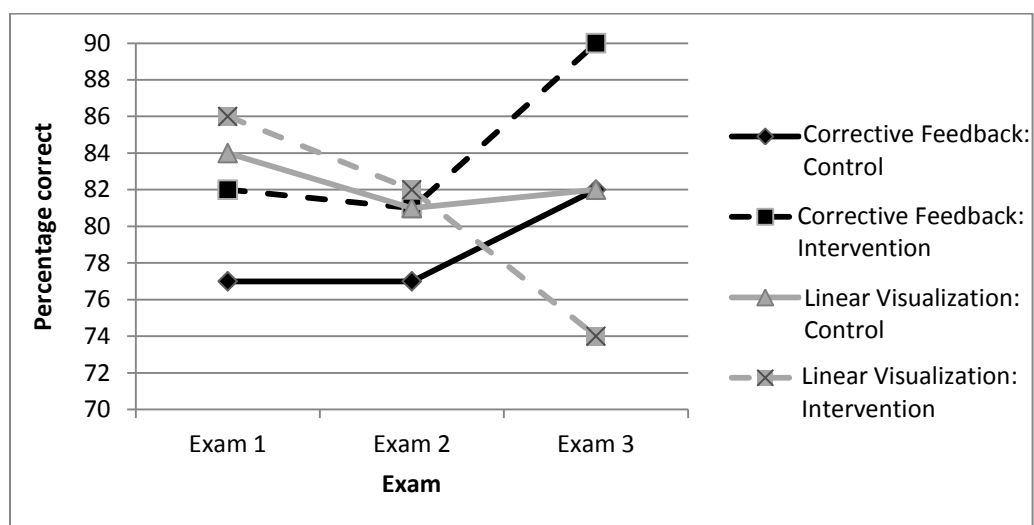


Figure 12. Comparison of control and intervention performance on exams by teacher

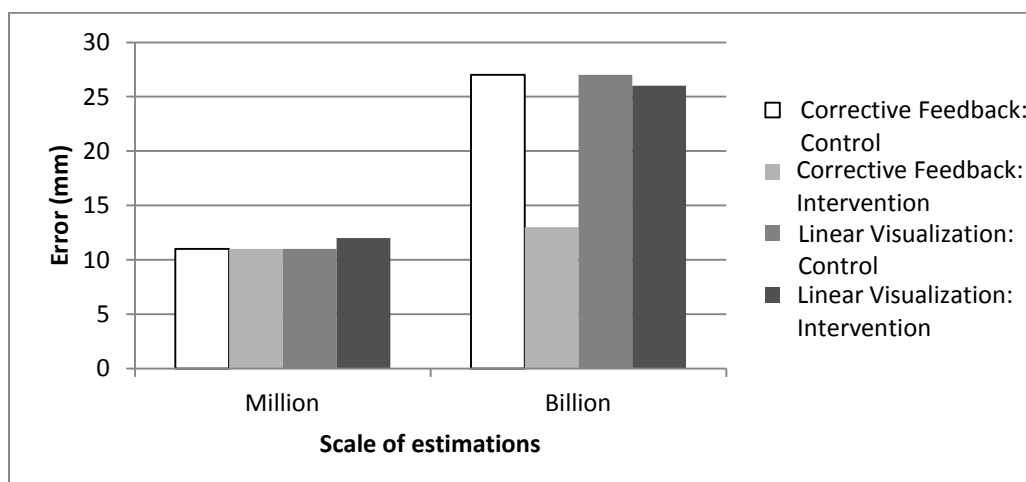


Figure 13. Comparison of control and intervention variance from correct answer on line estimation tasks by teacher

## Discussion

A main finding of the current study was, with the correct administration, the corrective feedback activity was associated with higher grades, and more linear representations of temporal and abstract (numeric) magnitudes. The corrective feedback activity was correctly administered in the corrective feedback class. Students from the corrective feedback class, who received the intervention, had higher exam scores, were more accurate on the GCI item measuring geologic time representation, and were more accurate on abstract line estimations than those students who received normal class instruction.

However, it is important to note, that students from the intervention condition had significantly higher grades on the pretest measures than the control condition. While this superior performance on pretest measures were statistically controlled for, higher marks on pretest measures may suggest the intervention group is comprised of a different population than in the control group. To clarify, students in the intervention condition may learn differently than the control group; in which case students from the different conditions may have learned from the corrective feedback activity differentially. Thus, these findings should be considered preliminary, and future research should examine more similar student samples.

The altered administration of the intervention (linear visualization) is actually serendipitous and informative. If linear visualization was successful at improving grades and developing linear representations of temporal and abstract magnitude, it would suggest that providing slide one, which pairs the Geologic Time Scale with a linear scale, is enough for students to make the alignment. If linear visualization failed to improve

grades or develop a linear representation of temporal or abstract magnitude, it would suggest that providing students with the aligned time lines is not enough for students to align the two scales. Further, unimproved grades, etc would highlight the advantages of the clicker response system in engaging the students to make the appropriate alignments.

In the current study, linear visualization alone actually led to lower exam scores. Linear visualization included only the presentation of the corrective feedback slide (see Figure 11). Students from the linear visualization class had significantly lower exam scores than those students who received normal class instruction. It is possible that this alternative administration of the corrective feedback activity interferes with understanding. Research on multiple external representations suggests that failure to align different representations can interfere with understanding (Ainsworth, Bibby, & Wood, 2002). Aligning different representations can be difficult for learners (de Jong, et al., 1998), and ability to make such alignments is characteristic of expert understanding (Kozma, Chin, Russel, & Marx, 2000). It is possible that because the students were not engaged in aligning the two time lines on slide one (see Figure 11), they may have failed to make the alignment altogether. Consequently, the students may not have understood the relationship between the two time lines, and having multiple presentations of the same information may have interfered with overall understanding.

Alternative accounts for why corrective feedback was associated with higher exam scores in one class, and linear visualization was associated with lower exam scores in a different class are not immediately apparent. There may be teacher-based differences between the corrective feedback and linear visualization classes. For example, the teachers may have presented the content outside of the corrective feedback activity

differently, or the teachers may have different levels of “buy-in” for the intervention itself. However, given the teachers’ willingness to participate in this study, and their collaboration developing similar curriculum, it is not clear that either of these teacher-based possibilities is the cause for the lower exam scores for students who received linear visualization compared with normal class instruction.

There may also be class-level differences. The corrective feedback and linear visualization classes were significantly different on one of the pretest measures (exam one) and not on the second pretest measure (exam two). Given the inconsistent finding, it is important not to over-interpret either pretest measure. If there was a difference between the corrective feedback class and the linear visualization class, there are many possible causes. For example, there may be student-based differences for each class, such as the time the course was scheduled may attract students with certain qualities. There may also be teacher-based differences, such as differences in lecture style or content focus (despite the teachers working together to develop similar curriculum). It may also be possible the differences on exam one fall under the five percent of expected false positive when analyzing mean differences. However, there were different attendance rates in the classes, which may suggest real class differences. The corrective feedback class had a higher attendance rate (76%) than the linear visualization class (55%).

The success of the corrective feedback activity with the correct administration, and the lack of success of when incorrectly administered, highlights the importance of the science of learning. The development of learning materials, or alteration of existing learning materials, using reasonable and common sense-based thinking, may sometimes not work. For example, an expert instructor may not recognize the effortful step needed

to align scales, and, thus, may not provide the structure that would help students coordinate representation. Unfortunately, a non-scientific approach to pedagogy results in the inability to determine why some activities work and others do not. Additionally, when analogies mislead students' understanding of a concept, it makes misconceptions hard to identify and resolve (Brown & Salter, 2010; Duit, 1991). By systematically assessing individual components of the analogical principles used in teaching scale information, it is possible to develop the most effective and efficient learning materials, as well as identify what approaches do not work and why. The current study advocates for providing students with multiple opportunities to align magnitude relations on a linear scale. Further research is needed to identify what about the alterations to the intervention led to a decrease in exam scores.

There were no significant differences between students who completed the corrective feedback activity and those who received regular class instruction on the multiple choice items added to the exam for the corrective feedback and linear visualization classes. There were also no differences between conditions on accuracy for multiple choice items with categorical response options compared to items with numerical response options. To review, four possible explanations were identified in chapter 2: students may not use magnitude representation to estimate correct responses, the multiple choice questions may not assess understanding of magnitude information required to understand certain scientific phenomenon, there may be sufficient learning in the control condition to detect differences in learning with the intervention, or the study may be under powered to detect differences in these particular questions. The control condition in the current study was normal class instruction, where students explicitly

learned about the geologic events on which they were tested. Thus, the educational approaches in the conventional conditions in both the Transfer study and the current study may lead to participant learning about relevant scientific phenomenon. Because the corrective feedback study has more participants ( $\mu=65$  in each condition) than the Transfer study, the corrective feedback study is less likely to be underpowered; with 50 to 75 participants in each condition, and using the standard 80% power, an effect size (Cohen's  $d$ ) of .5 (medium) is required to detect differences between conditions. However, effect sizes for these multiple choice items were all fairly small (Cohen's  $d < .1$ ). While there are no differences between conditions on the individual questions, there are exam-level differences. This may suggest an accumulation of differences is required to identify performance differences across conditions.

There were two types of multiple choice questions: questions that required magnitude recall only and those that required magnitude recall plus an additional step of reasoning. More accurate representation of temporal magnitude (as measured by the GCI item) and more linear abstract magnitude were not correlated with accuracy on recalling or reasoning about geologic magnitude information (as measured by the multiple choice questions requiring recall and recall plus reasoning). This is inconsistent with findings from the previous study (Transfer study), where temporal magnitude accuracy was associated with accuracy on items requiring magnitude recall plus reasoning and not items requiring magnitude recall only. Difference between findings from the two studies may be due to the measurement of temporal magnitude.

The Transfer study used a temporal line estimation task, which is a more sensitive measure of temporal magnitude representation than the GCI timeline item used in the corrective feedback study. Thus, the GCI timeline item may not be sensitive enough to detect relationships that are present between temporal magnitude representation and answering multiple choice items that require magnitude recall plus reasoning.

Two scales were developed for abstract (numeric) magnitude representations at the billion and million scales. All four abstract magnitude estimations at the billion scale were highly correlated and had strong internal consistency, suggesting these items were all measuring a single construct. Three of the four abstract magnitude estimations at the million scale were highly correlated and had strong internal consistency. The abstract magnitude estimation of 251 million out of 542 million was weakly correlated with the other estimations. A weak correlation is due to the increased accuracy all participants had on this item compared with the other abstract magnitude estimations. Because everyone does fairly well on this item, performance on this item does not strongly predict performance on other items. I hypothesize the increased accuracy when estimating 251 million out of 542 million is because 251 million is close to mid-point. Estimations near salient category boundaries are more accurate than those in between (e.g., Huttenlocher, et al., 1988; Haun, Allen, & Wedell, 2005), and even young children have been shown to divide one-dimensional spaces by the midpoint (Sandberg, Huttenlocher, & Newcombe, 1996). Increased accuracy due to proximity and saliency of midpoint is consistent with the hypothesis that large magnitudes are represented similarly to magnitudes at human scale: using a combination of categorical and metric information. This interpretation is consistent with previous work on large magnitude estimations, where magnitude

estimations near salient category boundaries were more accurate than those estimations not near boundaries (Resnick, et al., in prep). More research is needed to identify and characterize categories used during estimation. For example, adults have the tendency to naturally divide space into quadrants during recall (e.g., Crawford, et al., 2006; Huttenlocher, et al., 1988; Newcombe, et al., 1999). In order to examine if large abstract magnitudes are divided into quadrants, magnitudes near and far away from quadrants boundaries are required. Finally, performance on abstract magnitude estimation of 251 million out of 542 million also provides evidence that the participants understood the activity. While the other estimations may have been more difficult, because they were not near salient boundaries, participants were accurate when locating the midpoint.

There is evidence that learning a domain-specific magnitude (i.e., temporal magnitude) through the corrective feedback activity transfers to abstract magnitude representation. Students who received the corrective feedback activity had a more linear representation of abstract magnitude at the billion scale than those students who received normal class instruction. There were no significant differences for participants in both classes and both conditions on abstract magnitude estimations at the million scale; most participants were fairly accurate when making abstract magnitude estimations at the million scale. These differences in estimations at the million and billion scales, suggest that for most students abstract magnitude is represented differently at the million scale than the billion scale. Estimations at the million scale are linear and accurate; whereas estimations at the billion scale are overestimated but linear relative to each other. These findings are consistent with findings from the previous study (Transfer study), and are aligned with a segmented linear model of magnitude representation.

Abstract magnitude representation is not related to performance on exam or on individual multiple choice items. This suggests there is a dissociation between the representation of temporal and abstract magnitudes.

When students are not engaged in aligning magnitude scales (e.g., through use of the clicker response system), many students fail to spontaneously align the Geologic Time Scale with a linear scale. Students who were given linear visualizations were not significantly different from those who received normal class instruction on estimations of abstract magnitudes at the billion scale. That there were no differences between the linear visualization and control condition at the billion scale is inconsistent with findings from the previous study (Transfer study) and findings from the corrective feedback class. This suggests students from the linear visualization class were not engaged in making the alignments between the compressed image of the Geologic Time Scale and the linear time line.

## Chapter 4

### General Discussion

The Transfer study and Corrective Feedback study examined the way people naturally represent and reason about large magnitudes through the development of two science of learning tools: the hierarchical alignment activity and the corrective feedback activity. The hierarchical alignment activity utilizes the following analogical principles: hierarchical alignment, progressive alignment, structural alignment, and multiple opportunities to make analogies. The Transfer study examines the effectiveness of the principle of hierarchical alignment by contrasting the hierarchical alignment activity with a conventional activity that uses all the principles described above except for hierarchical alignment. The corrective feedback activity is based on the same analogical principles used in the transfer study, except it provides corrective feedback instead of progressive and hierarchical alignment. Taken together, the Transfer study and the Corrective feedback study offer an initial foray into parsing the individual principles of analogical reasoning useful for learning about scale information.

Both the hierarchical alignment activity and the corrective feedback activity were successful in developing a linear representation of domain specific magnitude. Participants who completed the hierarchical alignment activity were more accurate on the line estimation tasks (adapted from the Geologic Concept Inventory (GCI) item using time lines) for temporal and spatial magnitude estimations than participants who completed the conventional condition. Participants who completed the corrective feedback activity were more accurate on the original time line item from the GCI than participants who received normal class instruction. That both activities were successful in

developing linear representations of geologic time (and for the Transfer study, astronomical distances), suggests having multiple opportunities to make analogies through structural alignment are key components in developing analogies for learning scale information. However, because the hierarchical alignment activity was contrasted with an activity consisting of all the same analogical reasoning principles except for hierarchical alignment, it suggests there is an additive benefit of including hierarchical alignment in analogies for learning about scale information. Corrective feedback may also be a useful strategy in learning about scale information. More research is required to assess the individual contributions of each analogical principle.

In examining the effectiveness of the hierarchical alignment activity and the corrective feedback activity, the current studies also assessed large temporal, spatial, and abstract (numeric) magnitude representation. Both activities were based on the hypothesis that magnitudes at scales outside human perception are represented and reasoned about the same way as magnitudes at human scales. Magnitude information at human scales may be stored as a hierarchical combination of metric and categorical information (e.g., Crawford, et al., 2006; Huttenlocher, et al., 1988; Newcombe, et al., 1999; Zacks & Shipley, 2008). People may use category boundaries to help make estimations in lieu of precise metric information. Variation in estimation, therefore, occurs because of imprecision of category boundaries (Shipley & Zacks, 2008; Zacks & Tversky, 2001). The current studies, therefore, provided salient category boundaries to develop a more linear representation. Thus, the effectiveness of the hierarchical alignment activity and the corrective feedback activity supports the hypothesis that large magnitudes are represented the same way human scale magnitudes are represented.

While both interventions provided salient category boundaries by identifying magnitude relations, it is worth noting the interventions did so in slightly different ways. In the hierarchical alignment activity, participants aligned increasing amounts of magnitude, including important geologic divisions, onto a linear scale, and then located all previous magnitudes relative to the current magnitude. At each temporal scale, the previous magnitudes would make up different portions of the current scale. The nature of the hierarchical alignment principle is the practice of working with different magnitude relations at different scales. In the corrective feedback activity, students aligned important geologic divisions onto a linear scale. While the students still worked with different magnitude relations, all the magnitude relations were at the same scale (4.6 billion years). Because the corrective feedback activity used the time line item from the GCI to assess temporal magnitude representation, and the Transfer study used line estimation task, it is not possible to compare linearity of temporal magnitude representation. However, error in abstract magnitude for both studies is similar, suggesting the two techniques are equally effective in developing a linear representation of (at least) abstract magnitude. Future research may look into the saliency of category boundaries provided in the current studies, and if there are more effective category boundaries that could be provided to develop a linear representation of magnitude. Our lab is currently running a study to see if simply providing midpoint for temporal and abstract magnitudes is enough to foster more linear estimations. For example, for temporal estimations, does providing participants with the information “halfway between Earth forming and present day, there were only single cell bacteria on Earth” aide in the estimation of other geologic events (e.g., “when did fish appear”).

That the hierarchical alignment activity and the corrective feedback activity were successful at developing linear representations of magnitude across domains (temporal, spatial, and abstract magnitude) suggests that magnitude, irrespective of domain, is represented in similar ways. Similarity of magnitude representation across domains is consistent with previous research suggesting a generalized mapping of more/less relations across dimensions; such as time, space, number, and size (Brannon & Roitman, 2003; Lourenco & Longo, 2010; Walsh, 2003). However, participants performed differently on the temporal, spatial, and abstract magnitude measures. Participants in the Transfer study were the most accurate estimating abstract magnitudes, then spatial magnitudes, and the least accurate estimating temporal magnitudes. This could provide evidence against a generalized system of magnitude, consistent with findings of an asymmetrical relationship between spatial and temporal magnitudes (Agrillo, et al., 2010). Alternatively, it may be the case that temporal, spatial, and abstract magnitudes are all represented in a similar way, but preexisting knowledge (and misconceptions) bias the subjective categories people use to make estimations. For example, geologic time is often neglected in the classroom (Dodick, 2007; Trend, 2001) and learning about the solar system is commonplace. It seems likely participants had more knowledge of the solar system than geologic time, which is consistent with participants being better at estimating spatial magnitudes compared to temporal magnitudes. More research is needed to further identify and characterize categories used in the representation of geologic time and astronomical distances. Future research should examine unfamiliar scales, both in content and magnitude. Such a study might use an unfamiliar solar system, which would have a different temporal history and celestial objects.

The hierarchical alignment activity and the corrective feedback activity both develop a linear representation of domain specific and abstract (numeric) magnitude. This suggests that people use category boundaries when estimating both domain specific and abstract magnitudes. While there are some studies that look at the subjective categorization of numbers (Laski & Siegler, 2007; Mix, Huttenlocher, & Levine, 2002; Siegler & Robinson, 1982), no previous work has mapped the Category Adjustment Model (CAM) onto abstract magnitude representation.

People represent magnitude along a mental number line, with compressive effects on estimation of larger magnitudes (e.g., Booth & Siegler, 2008; Dehaene & Marques, 2002; Dehaene, et al., 2008; Opfer & Siegler, 2007; Siegler & Opfer, 2003). The nature of this representation is currently debated. Findings from the current studies are consistent with a segmented linear model. The segmented linear model suggests separate linear functions for sets of magnitudes (e.g., ‘familiar’ versus ‘unfamiliar’) when estimated magnitude is plotted against actual magnitude (Ebersbach, et al., 2008; Landy, et al., 2012). In both the Transfer study and Corrective Feedback study, participant error in estimation followed two separate linear functions for magnitude estimations at the million and billion scales. Participants were fairly accurate on magnitude estimations on the million scale, and overestimated magnitudes on the billion scale; however, these errors formed a linear pattern. Research including more magnitude estimations across scales (Opfer and Siegler (2003) recommend at least 24 magnitude estimations) is needed to more precisely characterize slope functions.

The CAM could serve as a unifying model for currently competing theories of magnitude representation. Two competing models of magnitude representation include a logarithmic-to-linear shift (Siegler & Opfer, 2003) and a power function with anchor points (Barth & Paladino, 2011). The CAM may, in some ways, unite these two competing models; category boundaries could serve as distinct anchor points, with adults possessing more precise categories (at the individual numbers level) compared with children. A third model is the segmented linear model (Ebersbach, et al., 2008; Landy, et al., 2012). As previously discussed, the findings from the current studies are consistent with the segmented linear model. The CAM may unite the segmented linear model with logarithmic and power models; whereas young children may have many unfamiliar numbers that are part of one “big” or “unfamiliar” category, adults may possess counting strategies for numbers within “unfamiliar” scales. To clarify, adults may over estimate where the billion scale is located on a number line that represents 10 trillion (Landy, et al., 2012); however, knowledge of the ten scale may allow for individual magnitude estimations at the tens of billions scale (e.g., 10 billion, 20 billion, etc) to be correct relative to other estimations at the tens of billions scale. More extensive research is needed to identify types of categories used in scale representation to see if a CAM can predict the changing pattern of bias in number line estimations that occurs with development.

Number line estimation tasks are useful tools to examine magnitude representation. Number line estimation tasks were used to measure abstract magnitude representation in the Transfer study and the Corrective Feedback study, and domain specific magnitude representation in the Transfer study. Magnitude representation may

involve knowledge and accurate cognitive representations of size (or magnitude) (Barth & Paladino, 2011), relevant measurement systems (e.g., miles, light years, pounds, etc), ratios and proportional reasoning (Jones & Taylor, 2009), and properties of the given scale. In both studies, the line estimation tasks had strong internal reliability and were highly correlated, which suggests the line estimations are measuring a single construct. There are also differences in performance across groups, domains, and scales. Training using the hierarchical alignment activity and the corrective feedback activity create more linear representations of magnitude compared with conventional training, which suggests the line estimations are sensitive enough to pick up on group differences.

There is also evidence that temporal, spatial, and abstract magnitude estimation tasks are not all measuring the same ability. Correlations are low, if not significant, between temporal, spatial, and abstract magnitude estimations. Temporal, spatial, and abstract magnitudes also have low internal reliability, suggesting there is more than one construct. Participants are most accurate on abstract magnitude estimations, then spatial estimations, and least accurate on temporal estimations. Participants are more accurate when making estimations of abstract magnitude on the million scale compared with the billion scale. The hierarchical alignment activity can develop more linear representations of spatial magnitude and not develop more linear representations of temporal magnitude. That participants can perform differently on different types of estimations (e.g., temporal, spatial, and abstract) suggests these line estimation tasks are assessing different abilities.

Thus, line estimation tasks can be useful tools in future research into magnitude representation. Development of magnitude representation scales using number line estimation tasks may include using different magnitudes/events/objects that elicit more variation in performance.

There are clear pedagogical implications to the current studies, as both studies aim to develop science of learning tools. The science of learning is an interdisciplinary field that aims to scientifically examine the way in which people naturally reason and learn, and to leverage those findings into effective teaching techniques. The Transfer study and Corrective Feedback study aim to examine the way in which people naturally reason about large magnitudes to develop an effective analogy for learning familiar to unfamiliar magnitudes. In developing an effective analogy, the current studies also examine principles of analogical reasoning. The corrective feedback activity is effective in developing a linear representation of geologic time, which translates to higher exam grades (a byproduct educators are sure to be interested in). As demonstrated by the hierarchical alignment activity, there is an additive effect of including hierarchical alignment in analogies teaching about geologic time. The hierarchical alignment activity takes longer to complete than the corrective feedback activity. Because of the scientific approach to parsing out the importance of each analogical principle, educators are informed on how to choose an analogy that suits their needs. In the Corrective Feedback study, the interference of learning seen in the linear visualization class highlights the importance of using tested analogical reasoning principles as prescribed.

Educators in any discipline that teach concepts requiring scales outside of human perception (e.g., evolution, global warming, astronomical distances, population growth,

governmental budgets, mathematics, and so on) would be recommended to teach the relevant scale information using analogical reasoning principles of hierarchical and progressive alignment. Students should be provided with the opportunity to practice mapping different magnitudes relations at different scales. Scales should be progressively aligned, starting with very familiar scales and moving to increasingly unfamiliar scales. Students should be provided with as many intermediate analogical steps required to make the alignment. Educators should be careful not to underestimate the amount of scaffolding required for students to make alignments. Analogies should be structurally aligned; meaning, all elements of the analogy should be the same except for the principle the educator is trying to teach. Further, the use of hierarchical and progressive alignment principles in different classes (e.g., math, geoscience, astronomy) may serve to help unify scale curriculum, as “size and scale” were identified as a fundamental and a unifying theme in science education (AAAS, 1993; Achieve, Inc., 2013; NRC, 2011).

## REFERENCES

- Achieve, Inc. (2013). Next Generation Science Standards (Crosscutting Concepts).  
Achieve, Inc.
- Agrillo, C., Ranpura, A., & Butterworth, B. (2010). Time and numerosity estimation are independent: Behavioral evidence for two different systems using a conflict paradigm. *Cognitive Neuroscience*
- Ainsworth, S.E., Bibby, P.A. & Wood, D.J. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Science, 11*(1), 25-62.
- American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. New York: Oxford University Press.
- Barghaus, K. & Porter, A. C. (2010, April). *Building aligned assessments for middle school science teachers and students*. Paper presented at the annual meeting of the American Educational Research Association., Denver, CO.
- Barth, H., & Paladino, A.M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science 14, 125-135*.
- Block, R. A. (1990). Models of psychological time. In R. A. Block (Editor), *Cognitive models of psychological time*. Lawrence Erlbaum: Hillsdale, New Jersey.
- Booth, J. & Siegler, R. (2008). Numerical Magnitude Representations Influence Arithmetic Learning. *Child Development, 79*(4), 1016-1031
- Boroditsky, L. (2001). Does language shape thought?: English and Mandarin speakers' conceptions of time. *Cognitive Psychology, 43*, 1-22

- Brannon, E.M. and Roitman, J. (2003) Non-verbal representations of time and number in non-human animals and human infants. In *Functional and Neural Mechanisms of Interval Timing* (Meck, W., ed.), pp. 143–182, CRC Press
- Brown & Salter. (2010). Analogies in science and science teaching. *Advanced Physiological Education*, 34, 167-169
- Casasanto & Boroditsky. (2008). Time in the mind: Using space to think about time. *Cognition*, 106, 579–593
- Catley & Novick. (2008). Digging Deep: Exploring College Students' Knowledge of Macroevolutionary Time. *Journal of Research in Science Teaching*, 46(3): 311-332
- Cheek, K.A. (2012). Students' understanding of large numbers as a key factor in their understanding of geologic time. *International Journal of Science and Mathematics Education*, 10(5), 1047-1069
- Chincotta, D., & Underwood, G. (1997). Bilingual memory span advantage for Arabic numerals over digit words. *British Journal of Psychology*, 88, 295-310.
- Clary, R. M. & Wandersee, J.H. (2009). Tried and True: How Old? Tested and Trouble free Ways to Convery Geologic Time. *ScienceScope*, 33(4), 62-66.
- Cowan, N., Baddeley, A.D., Elliott, E.M., & Norris, J. (2003). List composition and the word length effect in immediate recall: A comparison of localist and globalist assumptions. *Psychonomic Bulletin & Review*, 10, 74-79.
- Crannell, C.W. & Parrish, J.M. (1957). A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology: Interdisciplinary and Applied*, 44, 319-327.

- Crawford, E., Huttenlocher, J., & Hedges, L. V. (2006). Within-category feature correlations and Bayesian adjustment strategies. *Psychonomic Bulletin and Review*, *13*, 245-250.
- Dehaene. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4)
- Dehaene, S., Izard, V., Spelke, E., & Pica P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*
- Dehaene, S., & Marques, J. F. (2002). Cognitive neuroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quarterly Journal of Experimental Psychology*, *55*(3), 705–731.
- de Jong, T., Ainsworth, S., Dobson, M., van der Hulst, A., Levonen, J., Reimann, P., Sime, J., van Someren, M., Spada, H., & Swaak, J. (1998). Acquiring knowledge in science and math: the use of multiple representations in technology based learning environments. In M.W Van Someren, P. Reimann, H. Bozhimen, & T. de Jong (Eds.), *Learning with multiple representations* (pp 9–40). Amsterdam: Elsevier.
- Delgado, C., Stevens, S., Shin, N., Yunker, M., & Krajcik, J. (2007). *The Development of Students' Conceptions of Size*. A paper presented at the annual meeting of the National Association of Research in Science Teaching, New Orleans, LA.
- Dodick, J. (2007). Understanding evolutionary change within the framework of geological time. *McGill Journal of Education*, *42*(2), 245-264.
- Dodick, J., Orion, N. (2003). Cognitive factors affecting student understanding of geologic time. *Journal of Research in Science Teaching*, *40* (4), 415-442.

- Dodick, J.T., and Orion, N. (2006). Building an Understanding of Geological Time: A Cognitive Synthesis of the "Macro" and "Micro" Scales of Time, *in* Manduca, C.A. and Mogk, D.W., eds., *Earth and Mind: How Geologists Think and Learn about the Earth*, Geological Society of America, Special Paper 413, p. 77-93.
- Drane, D., Swarat, S., Hersam, M., Light, G., & Mason, T. (2008). An evaluation of the efficacy and transferability of a nanoscience module. *Journal of Nano Education*.
- Duit, R. (1991). On the role of analogies and metaphors in learning science. *Science Education*, 30, 1241–1257.
- Ebbinghaus, H. (1913). *Memory* (Trans. By H.A. Ruger and C.E. Bussenius), New York: Teachers College, Columbia University.
- Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2007). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9- year old children: Evidence for a segmented linear model. *Journal of Experimental Child Psychology*, 99, 1-17.
- Friedman, A., & Brown, N. R. (2000). Reasoning about geography. *Journal of Experimental Psychology: General*, 129, 193-219
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 203–222). Cambridge, MA: MIT Press.
- Gentner, D., Loewenstein, J., & Thompson L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393–408.

- Gentner & Namy. (2006). Analogical Processes in Language Learning. *Association for Psychological Science, 15*(6)
- Haun, D. B. M., Allen, G. L., & Wedell, D. H. (2005). Bias in spatial memory: a categorical endorsement. *Acta Psychologica, 118*, 149-170.
- Hermann, R. & Lewis, B. (2004). A formative assessment of geologic time for high school earth science students. *Journal of Geoscience Education, 52*, 231-235.
- Hofstadter, D.R. (1985). Metamagical themas: Questing for the essence of mind and pattern (Essay #6, "On number numbness", pp. 115–135). NY: Basic Books.
- Huart, J., Corneille, O., & Becquart, E. (2005). Face-based categorization, context-based categorization, and distortions in the recollection of gender ambiguous faces. *Journal of Experimental Social Psychology, 41*, 598-608.
- Huttenlocher, J., Hedges, L., & Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. *Psychological Review, 95*: 471–484.
- Huttenlocher, J. E., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*, 220-241.
- Izard, V. & Dehaene, S. (2008). Calibrating the mental number line. *Cognition, 106*, 1221-47
- Jones, M. G., Taylor, A., Minogue, J., Broadwell, B., Wiebe, E., & Carter, G. (2007). Understanding scale: Powers of ten. *Journal of Science Education and Technology, 16*(2), 191–202.
- Jones, M. G., Tretter, T., Taylor, A., & Oppewal, T. (2008). Experienced and novice teachers' concepts of spatial scale. *International Journal of Science Education, 30*(3), 409–429.

- Jones, M. G., Taylor, A. R., & Broadwell, B. (2009). Concepts of scale held by students with visual impairment. *Journal of Research in Science Teaching*, 46(5), 506–519.
- Jones, M. G., Tretter, T., Taylor, A., & Oppewal, T. (2008). Experienced and novice teachers' concepts of spatial scale. *International Journal of Science Education*, 30(3), 409–429.
- Jones, M. G., & Taylor, A. R. (2009). Developing a sense of scale: Looking backward. *Journal of Research in Science Teaching*, 46(4), 460–475.
- Kadosh & Walsh. (2009). Numerical representation in the parietal lobes: Abstract or not abstract? *Behavioral and Brain Sciences*, 32, 313–373
- Kotovskiy, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797–2822.
- Kozma, R., Chin, E., Russell, J., & Marx, N. (2000). The role of representations and tools in the chemistry laboratory and their implications for chemistry learning. *Journal of the Learning Sciences*, 9(3), 105–144.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Lamon, S. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. J. Harel Confrey (Ed.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 89–120). Albany, NY: State University of New York Press.

- Landy, D., Silbert, N., & Goldin, A. (2012). Getting off at the end of the line: the estimation of large numbers, *2012 Annual Meeting of the Cognitive Science Society Conference Proceedings*.
- Laski, E. & Siegler, R. (2007). Is 27 a Big Number? Correlational and Causal Connections Among Numerical Categorization, Number Line Estimation, and Numerical Magnitude Comparison. *Child Development, 78*(6), 1723-1743
- Lee, H.S., Liu, O.L., Price, A. & Kendall, A. (2011). College Students' Temporal-Magnitude Recognition Ability Associated with Durations of Scientific Changes. *Journal of Research in Science Teaching, 48*(3), 317-335.
- Libarkin, J.C., Anderson, S.W., Dahl, J., Beilfuss, M., & Boone, W. (2005). Qualitative Analysis of College Students' Ideas about the Earth: Interviews and Open-Ended Questionnaires. *Journal of Geoscience Education, 53*(1), 17-26
- Libarkin, J.C., Kurdziel, J.P. & Anderson, S.W. (2007). College student conceptions of geological time and the disconnect between ordering and scale. *Journal of Geoscience Education, 55*, 413-422.
- Lourenco & Longo, (2010). General Magnitude Representation in Human Infants. *Psychological Science*
- Mandler, G. (1967). "Organization and memory". In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory, 1*: 328-372. New York: Academic Press.
- Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review 63*, 81-97.

- Miller & Brewer. (2010). Misconceptions of Astronomical Distances. *International Journal of Science Education*, 32(12).
- Mix, K., Huttenlocher, J., & Levine, S.C. (2002). Multiple Cues for Quantification in infancy: Is number one of them? *Psychological Bulletin*, 128(2), 278-294.
- National Research Council. (2011). *A Framework for K-12 Science Education*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, DBASSE. Washington,DC: The National Academies Press.
- Neter J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of American Statistical Association*, 59,18-55.
- Newcombe, N.S., Huttenlocher, J., Sandberg, E., Lie, E., & Johnson, S (1999). What do misestimations and asymmetries in spatial judgment indicate about spatial representation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 986-996.
- Opfer, J. (n.d.). Analyzing the number-line task: A tutorial. Retrieved on March 1<sup>st</sup>, 2013, from <http://www.psy.cmu.edu/~siegler/SiegOpfer03Tut.pdf>.
- Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, 55, 169-195
- Opfer, J. E., Siegler, R.S., & Young, C.J. (2011). The powers of noise-fitting: Reply to Barth and Paladino. *Developmental Science*, 14, 1194 - 1204.
- Parcell, W. & Parcell, L (2009). Evaluating and communicating geologic reasoning with semiotics and certainty estimation. *Journal of Geoscience Education*, 57(5), 379-389.

- Park, J., Park, H., & Kwon, O. (2010). Characterizing Proportional Reasoning of Middle School Students. *The SNU Journal of Education Research*, 19, 119-144
- Petcovic & Ruhf. (2008). Geoscience Conceptual Knowledge of Preservice Elementary Teachers: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, 56(3), 251-260.
- Resnick, I., Shipley, T.F., Newcombe, N., Massey, C., Wills, T. (2011, October). *Progressive Alignment of Geologic Time*. Talk presented at the 2011 Geological Society of America Annual Meeting, Minneapolis, MN.
- Resnick, I., Shipley, T., Newcombe, N., Massey, C., Wills, T. (2012). Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning, *2012 Annual Meeting of the Cognitive Science Society Conference Proceedings*.
- Resnick, I., Atit, K., Shipley, T.F. (in press). Teaching Geologic Time with Geologic Events. Invited commentary on "The Significance of Geological Time: Cultural, Educational, and Economic Frameworks" by Cinzia Cervato and Robert Frodeman. Geological Society of America Special Publication.
- Richardson, R.M., 2000, Geologic time (clothes) line, *Journal of Geoscience Education*, 48
- Rips, L. (2012). How Many Is a Zillion? Sources of Number Distortion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*
- Roberson, D., Damjanovic, L. & Pilling, M (2007) Categorical Perception of Facial Expressions: Evidence for a 'Category Adjustment' model. *Memory & Cognition*, 35, 1814-1829.

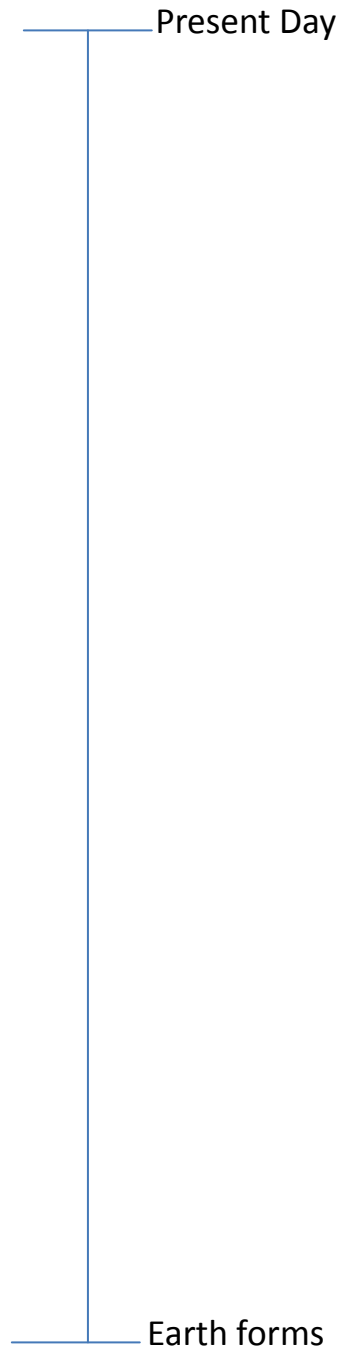
- Sandberg, E.H., Huttenlocher, J., & Newcombe, N. (1996). The development of hierarchical representation of two-dimensional space. *Child Development, 67*, 721-739.
- Semken, S., Dodick, J., Ben-David, O., Pineda, M.\*, Bueno Watts, N.\*, & Karlstrom, K. (2009). Timeline and time scale cognition experiments for a geological interpretative exhibit at Grand Canyon. Proceedings of the National Association for Research in Science Teaching, Garden Grove, California.
- Shipley, T. F. & Zacks, J. (2008). *Understanding Events: From Perception to Action*: New York, NY, Oxford University Press.
- Siegler, R. & Booth, J. (2004). Development of Numerical Estimation in Young Children. *Child Development 75*(2), 428-444.
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*.
- Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. *Advances in child development and behavior, 16*, 242-312.
- Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science, 18*, 449-455.
- Hawkins, D. (1978), Critical barriers to science learning, *Outlook, 29*
- Stevens, A., & Coupe, P. (1978). Distortions in judged spatial relations. *Cognitive Psychology, 10*, 422–437.
- Sullivan, J. & Barner, D. (2010). Mapping number words to approximate magnitudes: associative learning or structure mapping? 32nd Annual Meeting of the Cognitive Science Society.

- Swarat, S., Light, G., Park, E.-J., & Drane, D. (2011). A typology of undergraduate students' conceptions of size and scale: Identifying and characterizing conceptual variation. *Journal of Research in Science Teaching*.
- Thompson, C., & Opfer, J. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*.
- Trend, R.D. 1998, An investigation into understanding of geological time among 10- and 11-year-old children. *International Journal of Science Education*, 20(8), 973-988.
- Trend, R.D. (2000). Conception of geological time among primary teacher trainees, with reference to their engagement with geoscience, history and science. *International Journal of Science Education*, 22, 539–555.
- Trend, R.D. 2001, Deep Time Framework: a preliminary study of UK primary teachers' conceptions of geological time and perceptions of geoscience. *Journal of Research in Science Teaching*, 38(2), 191-221.
- Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006a). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43(3), 282–319.
- Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43(3), 282–319.
- Tretter, T. R., Jones, M. G., & Minogue, J. (2006b). Accuracy of scale conceptions in science: Mental maneuverings across many orders of spatial magnitude. *Journal of Research in Science Teaching*, 43(10), 1061-1085.

- Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006a). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43(3), 282-319.
- Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity *TRENDS in Cognitive Sciences*, 7(11)
- Wheeling Jesuit University. (2004). Geologic time activity. Copy right 1997-2004 by Wheeling Jesuit University/NASA-supported Classroom of the Future.
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127, 3-21.

APPENDIX A: ASSESSMENT OF TEMPORAL MAGNITUDE REPRESENTATION  
Questions presented below in the fixed order presentation seen by participants.

1. On the timeline below, please indicate the relative changes in life on Earth over time using the following events (not presented in any order): Humans appear, Life appears, Dinosaurs disappear, Dinosaurs appear



2. When did dinosaurs appear?

- A) Triassic    B) Jurassic    C) Archean    D) Proterozoic

3. When did dinosaurs disappear?

- A) 542 million years ago    B) 251 million years ago  
C) 65 million years ago    D) 2.6 million years ago

4. When did Pangea form?

- A) Archean    B) Paleozoic    C) Cretaceous    D) Cenozoic

5. When did Pangea start to break apart?

- A) 450 million years ago    B) 23 million years ago  
C) 200 million years ago    D) 2.7 billion years ago

6. How long were bacteria (Prokaryotes) the only life on Earth?

- A) 100 -100,000 years    B) 1-10 million years  
C) 500 -800 million years    D) 1-2 billion years

7. How far do you think continental plates move in a single year?

- A) A few inches (~50 millimeters)  
B) A few hundred feet (~50 meters)  
C) A few miles (~3 kilometers)  
D) We have no way of knowing  
E) Continental plates do not move

8. Evolutionary radiations occur on the scale of:

- A) 100 years or less  
B) Around 10 thousand years  
C) A few million years  
D) A hundred million years

9. If you could travel back in time to when the Earth first formed as a planet, what would the earth look like?

- A) The Earth would be mostly covered with water  
B) The Earth would be mostly covered with molten rock  
C) The Earth would be mostly covered with ice  
D) The Earth would be mostly covered with solid rock

10. Which of the following statements do you think best describes the relationship between people and dinosaurs?

- A) People and dinosaurs co-existed for about five thousand years
- B) People and dinosaurs co-existed for about five hundred thousand years
- C) Dinosaurs died out about five thousand years before people appeared on Earth
- D) Dinosaurs died out about five hundred thousand years before people appeared on Earth
- E) Dinosaurs died out about fifty million years before people appeared on Earth

11. According to the diagram below, which of the following statements is true?

- A. The Proterozoic Eon lasted much longer than the Phanerozoic Eon
- B. The Proterozoic Eon was much shorter than the Phanerozoic Eon
- C. The Jurassic Period ended 205 million years ago
- D. The Pre-Archean Eon is the most recent time span

| Eon         | Era   | Period  | Epoch   | Millions of Years Ago |
|-------------|---|---|---|-----------------------|
| Phanerozoic | Cenozoic  | Quaternary  | Holocene<br>Pleistocene                                 | 1.6                   |
|             |   | Tertiary  | Pliocene<br>Miocene<br>Oligocene<br>Eocene<br>Paleocene |                       |
|             | Mesozoic  | Cretaceous  | Late<br>Early   | 66                    |
|             |   | Jurassic  | Late<br>Middle<br>Early                                 | 138                   |
|             |   | Triassic  | Late<br>Middle<br>Early                                 | 205                   |
|             | Paleozoic                                       | Permian   | Late<br>Early   | 240                   |
|             |   | Pennsylvanian   | Late<br>Middle<br>Early                                 | 290                   |
|             |   | Mississippian   | Late<br>Early   | 330                   |
|             |   | Devonian  | Late<br>Middle<br>Early                                 | 360                   |
|             |   | Silurian  | Late<br>Middle<br>Early                                 | 410                   |
|             |   | Ordovician  | Late<br>Middle<br>Early                                 | 435                   |
|             |   | Cambrian  | Late<br>Middle<br>Early                                 | 500                   |
|             |   |   |   |                       |
|             | Proterozoic                                     | Late Proterozoic<br>Middle Proterozoic<br>Early Proterozoic |   |                       |
| Archean     | Late Archean<br>Middle Archean<br>Early Archean |   |   | 3,800?                |
| pre-Archean |   |   |   |                       |

12. When did dinosaurs appear?

- A) 230 million years ago      B) 3.5 billion years ago  
C) 398 million years ago      D) 2 billion years ago

13. When did dinosaurs disappear?

- A) Beginning of the Cambrian      B) End of the Permian  
C) End of the Cretaceous      D) Beginning of the Pleistocene

14. When did Pangea form?

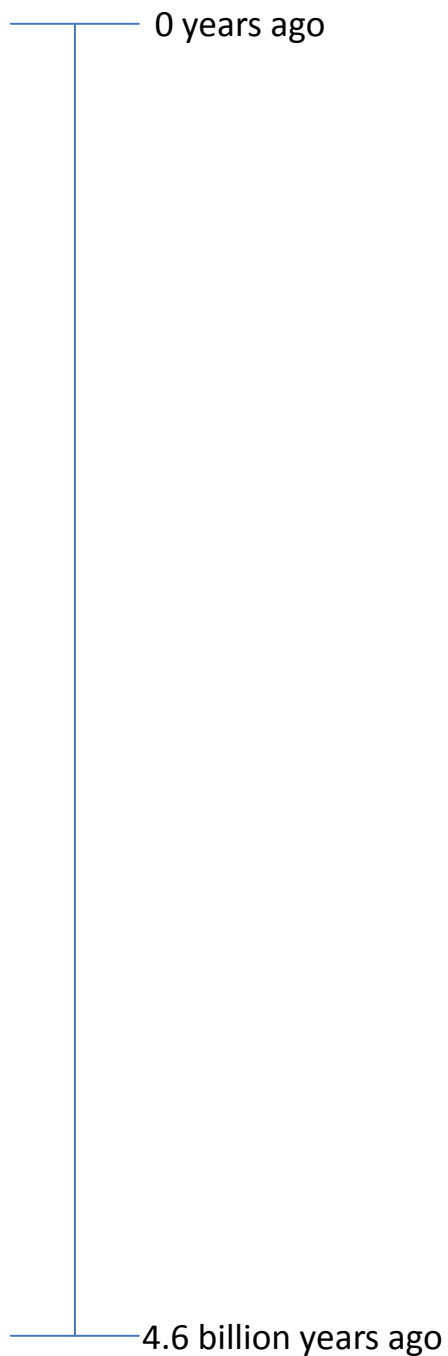
- A) 2.8 billion years ago      B) 300 million years ago  
C) 65 million years ago      D) 20 million years ago

15. When did Pangea start to break apart?

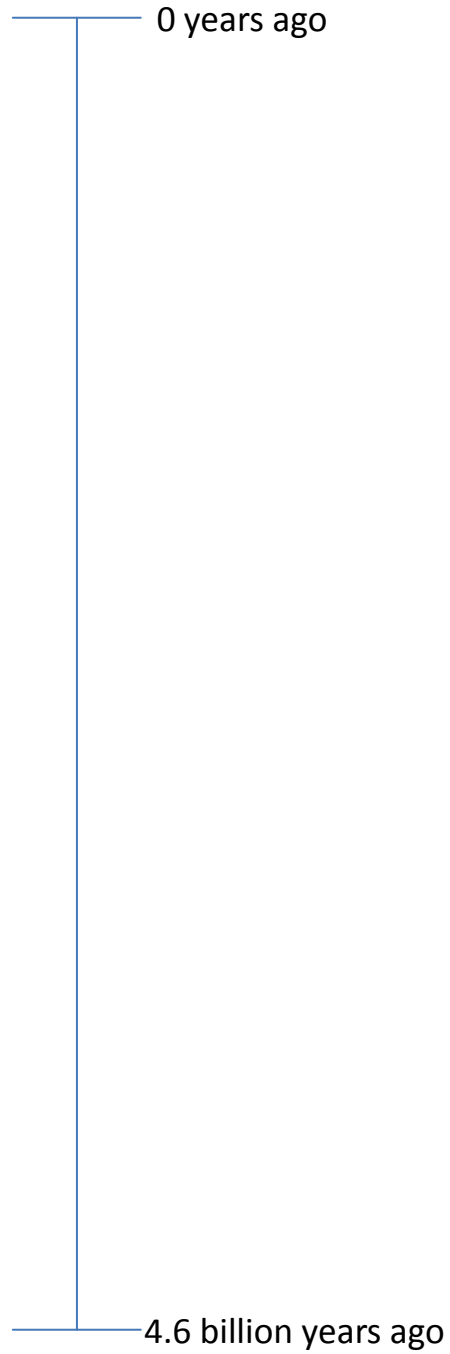
- A) Paleozoic      B) Cenozoic      C) Triassic      D) Archean

For the following 2 questions please use the following time line, which extends from 0 years ago to 4.6 billion years ago.

A) The Rocks of Sierra Nevada Mountains formed between 125 million years ago and 85 million years ago. Draw on the timeline provided when these rocks formed.

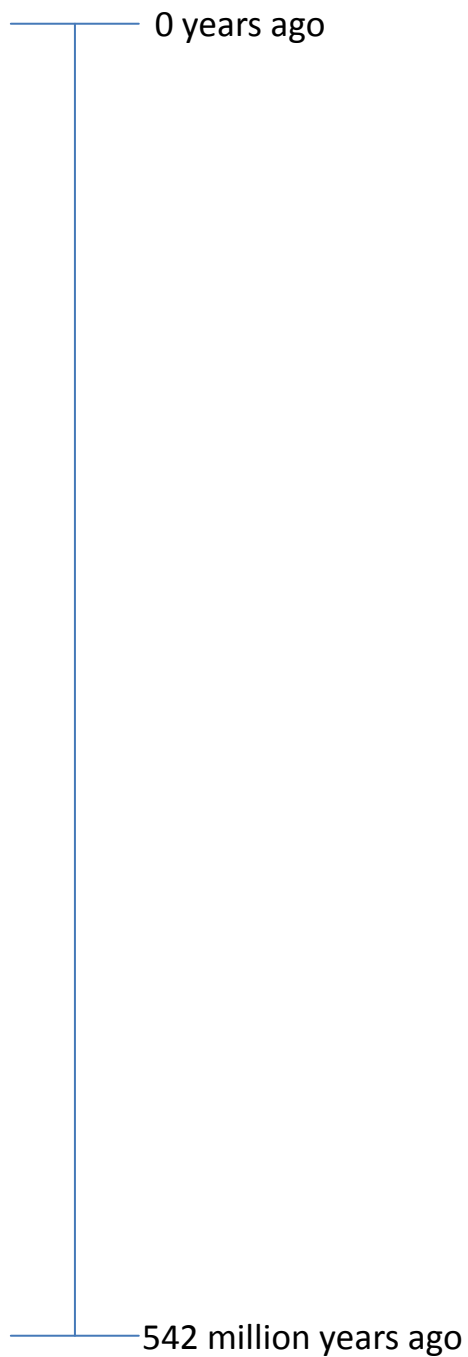


B) The Paleozoic began 542 million years ago and ended 251 million years ago. Please draw on the time line provided when the Paleozoic occurred.

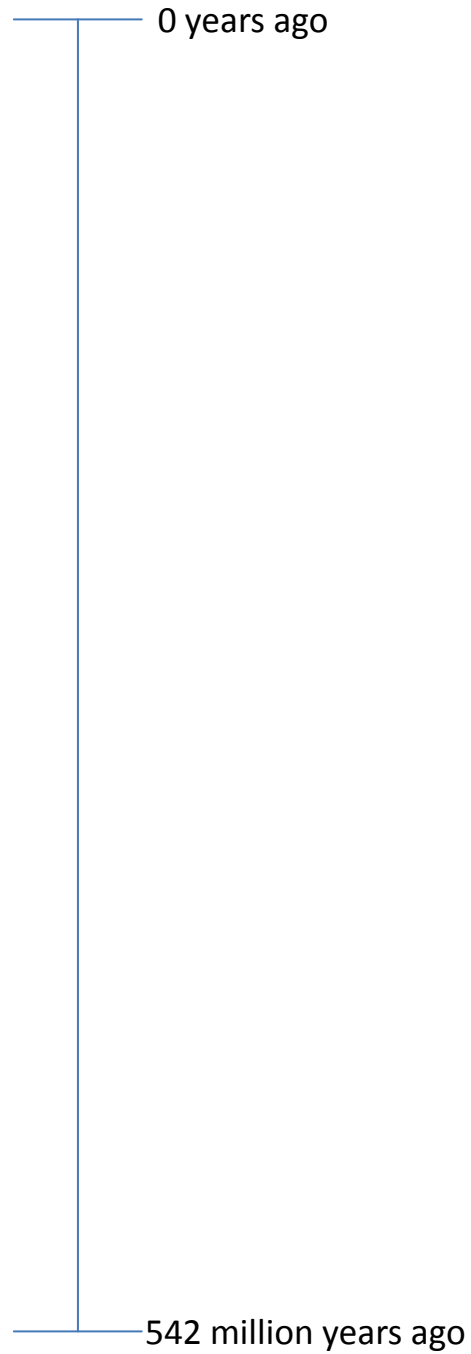


For the following 2 questions please use the following time line, which extends from 0 years ago to 542 million years ago.

A) The Laramide orogeny was a period of mountain building in western North America. The Laramide orogeny occurred between 70 million years ago and 35 million years ago. Draw on the timeline provided when these rocks formed.

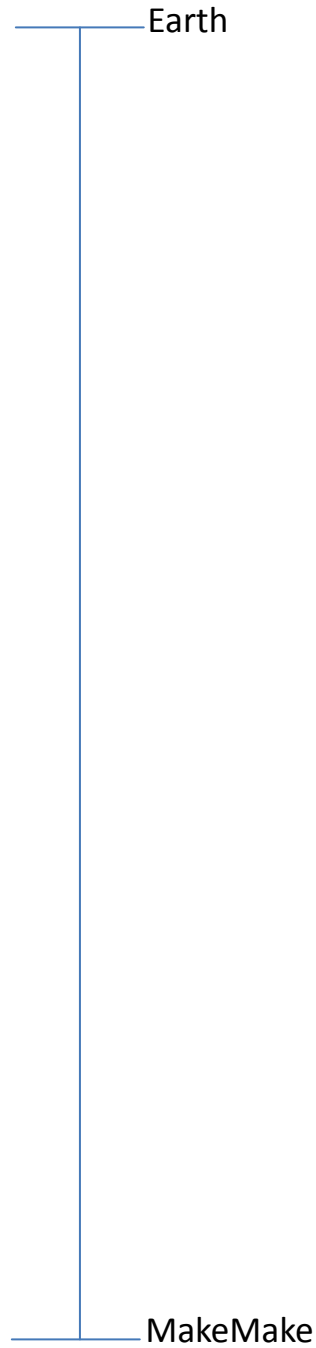


B) The Mesozoic began 251 million years ago and ended 65 million years ago. Please draw when the Mesozoic occurred.



APPENDIX B: ASSESSMENT OF SPATIAL MAGNITUDE REPRESENTATION  
Questions presented below in the fixed order presentation seen by participants.

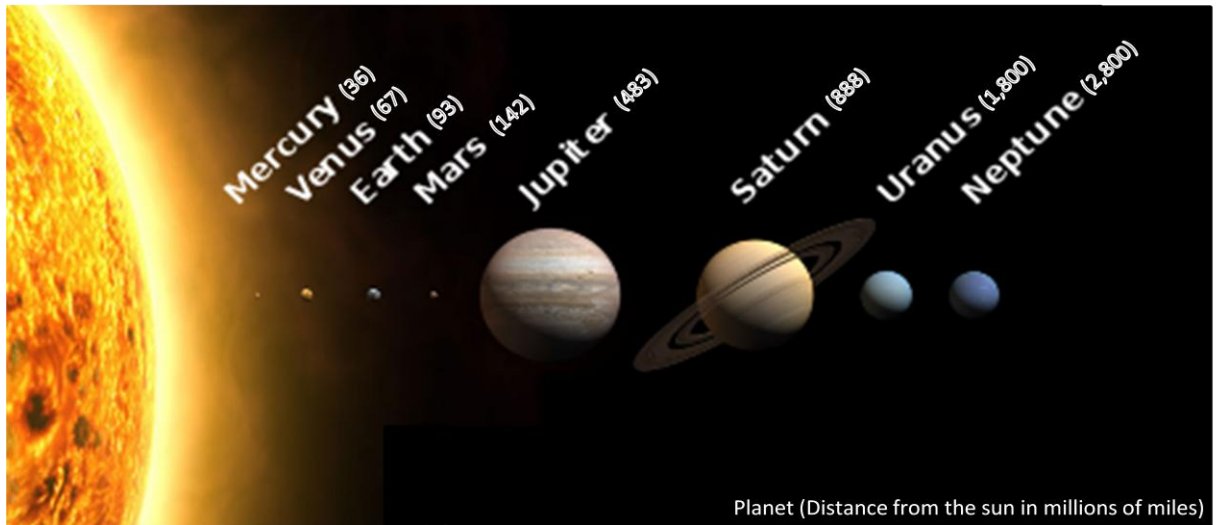
1. On the scale below, please indicate the locations of the following objects of the Solar System (not presented in any order): 3753 Cruithne, Pluto, Mercury, Mars



2. How far away from the Earth's surface is the international space station?  
A) 250 miles                      B) 6,000 miles  
C) 52 miles                         D) 11 miles
3. How far away from the Earth's surface is Mercury?  
A) Kuiper Belt                      B) Inner Van Allen Belt  
C) Inner solar system              D) Outer solar system
4. How far away from the Earth's surface is Saturn?  
A) 3,580,000,000 miles          B) 6,000 miles  
C) 57,000,000 miles              D) 777,000,000 miles
5. How far away from the Earth's surface is Jupiter?  
A) Inner solar system              B) Outer Solar System  
C) Kuiper Belt                        D) Inner Van Allen Belt
6. How far do you have to travel from Earth to reach the farthest planet in this Solar system?  
A) 100-1,000 miles                B) 1 – 10 million miles  
C) 500 – 800 million miles      D) 2-3 billion miles
7. Approximately how far do you think the Earth moves in a single year (Earth's orbit around the sun)?  
A) 585 million miles                B) 95 million miles  
C) 2 billion miles                    D) 600,000 miles
8. How long would it take to travel to Mars?  
A) less than 6 months              B) 6 months -1 year  
C) 2-3 years                          D) 3 – 10 years  
E) more than 10 years

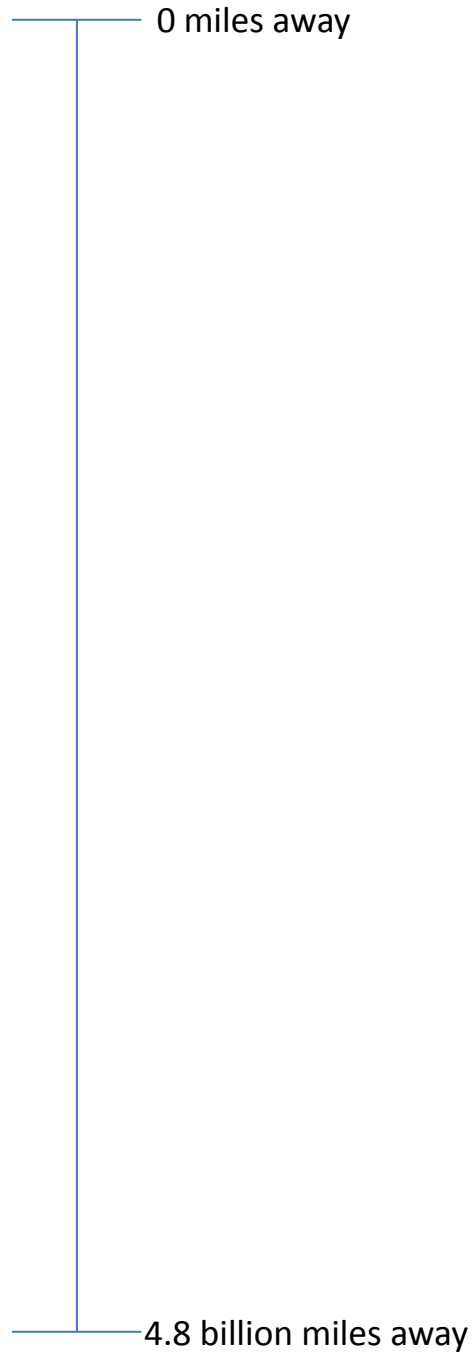
9. How far away from the Earth's surface is the international space station?  
A) Exosphere                      B) Inner Van Allen Belt  
C) Middle Atmosphere            D) Troposphere
10. How far away from the Earth's surface is Mercury?  
A) 3,580,000,000 miles          B) 6,000 miles  
C) 57,000,000 miles              D) 777,000,000 miles
11. How far away from the Earth's surface is Saturn?  
A) Kuiper Belt                      B) Inner Van Allen Belt  
C) Inner solar system              D) Outer solar system
12. How far away from the Earth's surface is Jupiter?  
A) 8,450,000 miles                B) 400,000,000 miles  
C) 2,600,000,000 miles          D) 6,000 miles

15. According to the diagram below, which of the following statements is true?
- A) Saturn is closer to Venus than it is to Neptune
  - B) Saturn is closer to Neptune than it is to Venus
  - C) Venus is the smallest planet
  - D) Eris is the closest object to the sun

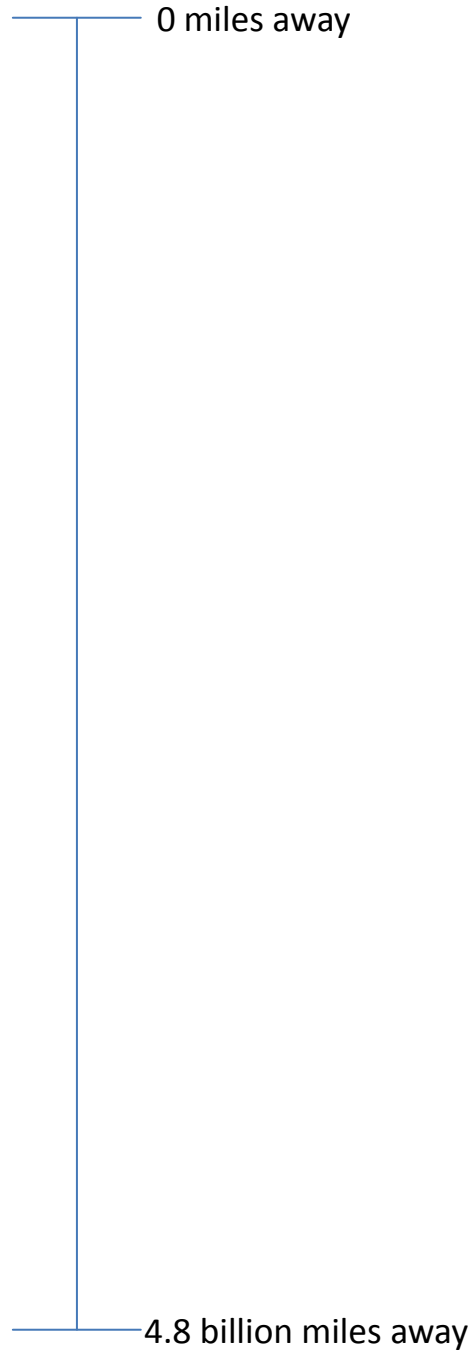


For the following 2 questions please use the following scale, which extends from Earth's surface (0 miles) to 4.8 billion miles away.

A) Venus is 26 million miles away from Earth. Please draw on the line provided where Venus is located.

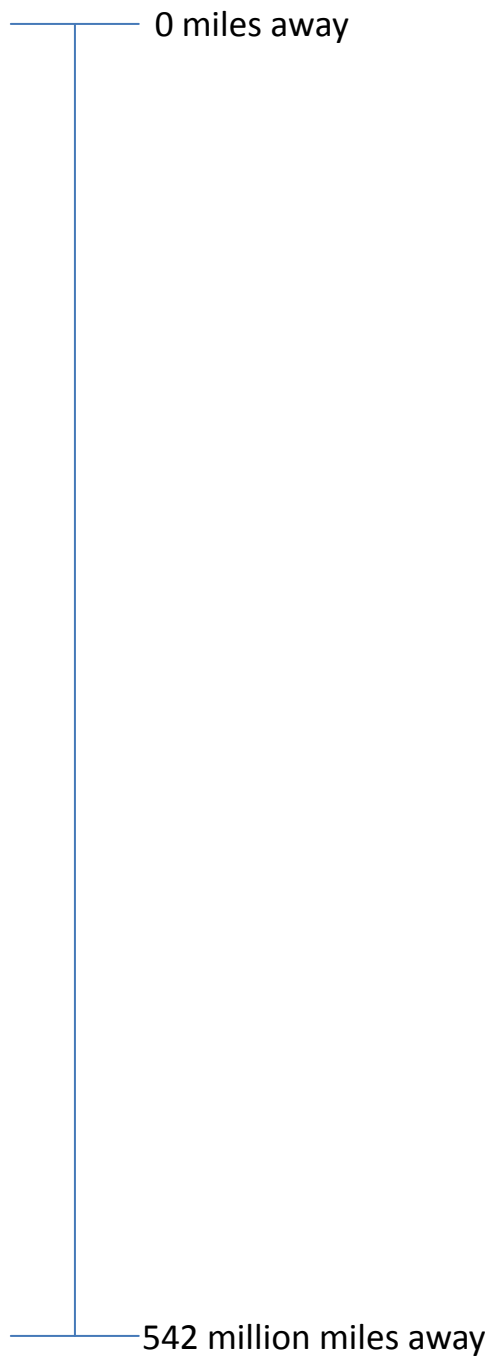


B) An asteroid belt formed approximately 215 million miles from Earth. Please draw on the line provided where the asteroid belt is located.



For the following 2 questions please use the following scale, which extends from the Earth's surface (0 miles away) to 542 million miles away.

A) At its closest, Mars is located approximately 35 million miles away from Earth. Draw on the line provided where Mars is located.



B) The protoplanet called Ceres is approximately 250 million miles from Earth. Please draw on the line provided where Ceres is located.

