JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 48, NO. 3, PP. 317–335 (2011)

College Students' Temporal-Magnitude Recognition Ability Associated With Durations of Scientific Changes

Hee-Sun Lee,[1,2] Ou Lydia Liu,[3] C. Aaron Price,[1] and Amber L.M. Kendall[1]

[1]*Education Department, Tufts University, Paige Hall Room 201, Medford, Massachusetts 02155*
[2]*Graduate School of Education, University of California, Berkeley, Berkeley, California*
[3]*Educational Testing Service, Princeton, New Jersey*

Received 9 April 2010; Accepted 28 October 2010

Abstract: The purpose of this study was to explore college students' recognition of temporal magnitudes associated with durations of scientific changes through construct validation of a 30-item instrument. We administered the instrument to 514 students from 10 higher education institutions in the United States. Among them, 419 students took the instrument at the end of science courses. The remaining 95 students took the instrument before and after a course on cosmic evolution, and additionally answered whether they knew, estimated, or randomly guessed at the temporal magnitude of each item in the instrument. We also collected exam scores for the cosmic evolution course students. Using descriptive statistics and a Rasch analysis, we investigated construct validity of the instrument in terms of psychological relevance, psychometric conformity, and instructional sensitivity. Results of this study indicate that (1) the temporal-magnitude recognition ability is a measurable construct, (2) extremely small duration items are significantly more difficult for students to recognize accurate temporal magnitudes than other duration range items, (3) direct knowledge of the magnitude contributes to the measurement of the construct, (4) the instrument is sensitive to instruction designed to improve the construct, and (5) the temporal-magnitude recognition ability is not significantly correlated with knowledge about the related scientific changes. J Res Sci Teach 48: 317–335, 2011
Keywords: temporal magnitude; construct validity; duration; scientific change; Rasch analysis

Change is everywhere in the universe (Chaisson, 2001). Much of science and mathematics is devoted to describing and explaining how change occurs (American Association for the Advancement of Science 1993). Change is generally described as ''becoming different'' (National Research Council 1996, p. 117) over time. Time concerns ''two different, though often related, notions: the first is 'interval' which means *duration* in time and the second is 'epoch' which means *location* in time'' (Roeckelein, 2000, p. 1). This duality makes the quantity expressed in number and unit to describe the temporal dimension of change ''both ordinal and cardinal: the order of events, or the ordinal succession of reference points, corresponds to the cardinal value, or duration of the intervals between these points'' (Piaget, 1969, p. 38). In the ordinal sense of time, some scientific changes can be sequenced, in the order they occurred, from the beginning of the Universe to the present time (McPhee, 1981; Trend, 2001). The temporal magnitude associated with the duration during which a change occurs also has a cardinal value and can be perceived, measured, or estimated. These durations of scientific changes range from less than a billionth of a second to billions of years.

Such vast differences in temporal magnitudes across scientific changes are hard to comprehend without technological aids or theoretical conceptualizations because we humans cannot directly experience changes which lie outside the range of 100 milliseconds (Ward, 1975) up to approximately 100 years. Yet, adequate

Additional Supporting Information may be found in the online version of this article.
Correspondence to: Hee-Sun Lee; E-mail: heesun.lee@tufts.edu
DOI 10.1002/tea.20401
Published online 9 December 2010 in Wiley Online Library (wileyonlinelibrary.com).

recognition of temporal magnitude is needed because ''some 'laws' of science ... are valid only within a certain range of circumstances'' (AAAS, 1993, p. 276). If students do not differentiate between temporal magnitudes of atomic or astronomical phenomena and those of everyday phenomena, they are likely to extend properties meaningful in everyday scales to those in realms of statistics, quantum physics, or relativity (Castellini et al., 2007). Likewise, if students are not aware of immense temporal magnitudes, the teaching of biological evolution or geological changes can be very difficult (Trend, 2002; Zen, 2001).

No prior research has directly addressed perceptions of, or conceptions about, durations across all temporal magnitudes in one single study (Dodick & Orion, 2003a). In theorizing students' recognition of temporal magnitudes from extremely short to long ranges, we synthesized five related research areas: (1) human duration perception between milliseconds and hours (for review, see Block, 1990; Roeckelein, 2000), (2) spatio-temporal reasoning (Boroditsky, 2000; Gibbon & Church, 1981; Torralbo, Santiago, & Lupianez, 2006), (3) number sense and number line (Dehaene, Izard, Spelke, & Pica, 2008; Walsh, 2003), (4) spatial scale conception from a billionth of a meter to a billion meters (Tretter, Jones, Andre, Negishi, & Minogue, 2006; Tretter, Jones, & Minogue, 2006), and (5) temporal-magnitude conception from the Big Bang to the emergence of modern humans (Catley & Novick, 2009; Dodick & Orion, 2003a, b; Trend, 1998, 2000, 2001). Studies in these areas described subjects' conceptions of spatial, temporal, and numerical magnitudes in individual items sampled to represent different magnitude ranges. The next step in research is to examine whether student responses to these individual items can be attributed to an overall underlying ability for scale conception. For example, is the same underlying cognition responsible for recognizing temporal magnitudes of the tidal cycle, the age of the Sun, and the shortest time required for a chemical reaction?

In this study, we explored college students' ability to recognize temporal magnitudes associated with durations of scientific changes by examining construct validity of a 30-item instrument designed to measure the ability. As recommended by Messick (1989), we used construct validity as a framework to interpret evidence from psychological, psychometric, and instructional sources. Research questions of this study are:

- How well do student responses to the items in the temporal-magnitude instrument match theory and findings in the literature on magnitude perception and scale conception?
- How well do the items in the instrument contribute to the measurement of the temporal-magnitude recognition construct?
- How are students' knowledge, estimation, and random guessing of temporal-magnitudes associated with the measurement of their ability on the construct?
- Is the instrument sensitive to instruction intended to improve recognition of extremely large temporal magnitudes?
- How is students' temporal-magnitude recognition ability correlated with their knowledge of the related scientific phenomena?

We first define terms related to quantity, magnitude, and scale. We then synthesize literature related to changes across temporal magnitudes, current standards on teaching temporal-magnitude recognition, and studies on perceptions of duration and conceptions of magnitude. We then describe research methods, including subjects, test design, and data collection and analysis. In the results section, we compare students' response patterns with findings from the literature on duration perception and scale conception, describe a measurement scale resulting from a Rasch analysis, and examine pre-posttest differences to determine instructional sensitivity. Finally, we discuss implications for science teaching and future research directions.

Quantity, Magnitude, and Scale

Quantity refers to the ''property of a phenomenon, body, or substance where the property has a magnitude that can be expressed as a number and a reference'' (Joint Committee for Guides in Metrology, 2008, p. 2). Mathematically, scale refers to a set of ordered quantities that represent a particular type of quantity. On a scale, we can compare and rank quantities of the same type according to their magnitude. The connection between number and measurement through scale is a topic which has been extensively discussed in philosophy and history of science (Campbell, 1920; Darrigol, 2003).

This study focuses on durations of scientific changes as numerically represented quantities with appropriate units of time. The International System of Units (SI) lists time as one of the *base* quantities that cannot be expressed in terms of other quantities and the *second* as a base unit for time (JCGM, 2008). Units such as the minute, hour, and day are ''officially considered non-SI units [but are] accepted for use with the International System'' (Whitelaw, 2007, p. 91). Numerical values that represent the various durations of scientific changes can be used to rank these durations on a common time scale according to their magnitude. For example, the average duration of a breath is about 4 seconds, and that of solar eclipse is about 360 seconds (6 minutes). Therefore, these two different scientific changes (a breath vs. a solar eclipse) can be compared and ranked. The duration of a breath is shorter than that of a solar eclipse based on their quantities (4 seconds vs. 360 seconds) according to magnitude (4 vs. 360) on the common time scale using seconds as a unit of reference.

Throughout this article, we use temporal magnitude to refer to a quantity related to duration expressed in number and unit (e.g., 50 seconds or 1,000 light years). Note that this time scale is different from the scale we developed to measure the temporal-magnitude recognition construct in this article (Michell, 1999). To distinguish, we use ''measurement scale'' for the temporal-magnitude recognition construct, as compared to time scale, throughout this article.

Scientific Changes Across Temporal Magnitudes

In nature, different patterns of change are observed. *Science for All Americans* (AAAS, 1990) identifies: ''(1) changes that are steady trends, (2) changes that occur in cycles, and (3) changes that are irregular'' (p. 175). *Trends* refer to overall changes with a predictable direction that a system, object, or organism goes through over time (AAAS, 1990). For example, once food enters the human body, digestion of the food begins; the direction of change in this digestion process is from whole food to component compounds, and is thus a trend. *Cycles* refer to changes that occur in the same system or object with defined periodicity. Some examples are seasons, the phases of the moon, and tides. Sometimes, *irregular changes* are observed in constituents of a much larger system. ''Most systems above the molecular scale involve the interactions of so many parts and forces and are so sensitive to tiny differences in conditions that their detailed behavior is unpredictable'' (AAAS, 1990, p. 173). Despite this complexity, at the system level some random and chaotic changes of constituents lead to more predictable, macroscopic changes which can be observed as *trends* or *cycles*. Evolution is another type of change that occurs over much longer time (Chaisson, 2006; Trend, 2001). *Evolutionary changes* are different from *trends* in the sense that they take much longer time to occur, involve multiple generations, and are unpredictable.

Physically, the durations of all scientific changes are bounded by two quantities: the smallest being one Planck time, 5.39×10^{-44} seconds according to quantum theory, and the largest being approximately 13.73 billion years, or the age of universe (Hinshaw, 2009). The bounds on durations that humans can experience are much closer together than this extreme range. For this study, we define the human-experience range from 100 milliseconds to a human's lifetime (approximately 100 years), because durations shorter than 100 milliseconds are perceived by humans as instantaneous (Ward, 1975). Therefore, we use ''extremely small'' temporal magnitudes to refer to shorter than 100 milliseconds and ''extremely large'' to longer than 100 years in this article.

The ability to recognize the temporal magnitudes of scientific changes relies on an individual student's knowledge of scientific change, sense of numbers, and familiarity with the time scale. The *National Science Education Standards* (NRC, 1996) consider change to be one of the unifying concepts that ''provide connections between and among traditional scientific disciplines'' (p. 115). *Benchmarks for Science Literacy* considers change and scale as two of the common themes ''that pervade science, mathematics, and technology and appear over and over again ... that transcend disciplinary boundaries and prove fruitful in explanation, in theory, in observation, and in design'' (AAAS, 1993, p. 261).

Benchmarks (AAAS, 1993) recommended that teaching about change should involve observing changes and identifying change patterns (grades K-5), using measurements to describe change (grades 6–8), understanding energy and entropy as driving forces of change (grades 9–12), and recognizing uncertainty involved in change (grades 9–12). Teaching scale focuses on attending to differences in physical

variables (grades K-2), recognizing extremes of the physical variables (grades 3–5), understanding scaling effects on system properties and complexity (grades 6–8), and representing extremely large and small numbers based on powers of 10 (grades 9–12) (AAAS, 1993). According to these recommendations, by the end of 12th grade, students should be able to recognize temporal magnitudes associated with scientific changes in physical, biological, and earth sciences using powers of 10 notations with appropriate time units.

Perceptions of Duration and Conceptions of Magnitude

Time is difficult, if not impossible, to define (Roeckelein, 2000), but ''the accurate measurement of time is central to an understanding of the laws and processes at work in the Universe'' (Whitelaw, 2007, p. 87). In science, time is considered the fourth dimension along with three space dimensions in describing scientific objects and processes. Unlike space dimensions, the time dimension (t) has an origin and flows continuously and unidirectionally everywhere in the known Universe (Chaisson, 2001, 2006). Due to these characteristics, the temporal aspects of scientific phenomena are often described in terms of relative position in time, succession (order), duration, and simultaneity (Roeckelein, 2000). Historically, time-related concepts have been defined and refined through advances in time measurement (Barnett, 1999), theory (Einstein, 2005; Newton, 1995), wave applications (Bracewell, 1999), and thermodynamic interpretations (Chaisson, 2001). In this review, we focus on durations of which temporal-magnitudes subjects of this study were asked to recognize.

Time can be characterized in three distinctive forms: absolute, physical, and psychological. Sir Isaac Newton distinguished absolute from relative (physical) time by saying that ''absolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly . . . [Physical time] . . . is any sensible and common external measure (precise or imprecise) of duration by means of motion, [and] such a measure—for example, an hour, a day, a year—is commonly used instead of true [absolute] time'' (Whitelaw, 2007, p. 88). For instance, periodic motions such as the Earth's revolution around the Sun have long been used to measure physical time. A more precise measure of time adopted by the International System of Units (SI) also has a physical reference, given that one second is defined as ''the duration of 9,192,631,770 periods of radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom'' at zero Kelvin (Whitelaw, 2007, p. 91).

In contrast, psychological time is perceived by an individual. Its complex nature is well recognized as numerous temporal constructs concerning succession, duration, and perspective have been identified in the field of psychology (Block, 1990). According to Piaget (1969), people's evaluation of duration is built on their grasp of simultaneity and succession of events and, by age 10 or 11, most children enter into the stage of ''operational construction of qualitative durations and the measurement of time'' (p. 69). However, this operational competency with qualitative duration does not guarantee that one can accurately estimate quantitative duration.

Quantitative judgments of duration are rooted directly in the changes people experience, and later, in the changes they recall (Fraisse, 1984) because ''changes serve as referents, or cues, to use in experiencing, remembering, and judging time'' (Block, 1990). The range of quantitative duration that people can directly experience is limited. Without external referents, subjects can accurately reproduce durations in the range from 500 milliseconds to a few seconds (Block, 1979). This range is called ''the indifference point,'' a ''time period that is, on the average, neither overestimated nor underestimated'' (Block, 1979, p. 185), though the exact numerical values for the range are not in agreement. People tend to underestimate the target durations longer than a few seconds (Eisler, 1976) and overestimate the target durations shorter than 100 milliseconds (Ward, 1975).

Zakay and Block (1997) distinguished experienced duration from remembered duration. To study experienced duration, subjects are directed to pay attention to time-in-passing during a target duration and later asked to estimate the target duration verbally or reproduce the duration by delimiting a matching time period manually. When engaged with more complex tasks during the target duration, subjects' experienced duration estimation becomes shorter in part because they cannot fully allocate their attention to estimating the target duration (Brown, 1985). According to Block (1990), remembered duration involves more complicated cognitive and biological mechanisms than experienced duration, and is an outcome of the interaction among:

''characteristics of the person who experiences the time, contents of the event, the person's activities during the time period, and the person's time-related behaviors and judgments'' (p. 31). This implies that remembered durations can be easily distorted, making accurate estimation of durations difficult, even when experienced first-hand.

However, direct application of these psychological studies to the recognition of durations related to scientific changes across all temporal magnitudes is limited because (1) they rarely address events shorter than a few milliseconds and longer than a few days, and (2) the events or stimuli used in these studies, such as controlled blinking lights or audio signals, are not often pertinent to science knowledge students need to learn in school.

In science education, estimation of large temporal-magnitudes has been investigated as event-based studies in the context of deep time (Dodick & Orion, 2003a). In event-based studies, students were provided with a set of geological events, such as appearance of the first mammals in pictures or words. Students were then asked to generate their own duration estimates for an event in a free response format (Catley & Novick, 2009), select one of the temporal-magnitude categories provided in a forced choice format (Trend, 2000, 2001), or determine the temporal order of the events (Trend, 1998, 2000, 2001). Trend (1998) argued that elementary school students aged 10–11 had a general awareness of some geological events without clear chronology and tended to put the geological events into two broad categories: extremely ancient and less ancient. With pre- and in-service primary school teachers, Trend (2001) found a more improved, but not perfect, chronology of the geological events with three distinct categories: extremely ancient, moderately ancient, or less ancient. Catley and Novick (2009) found that, when generating duration estimates in an open-ended format, college students showed large variations in their estimates of temporal magnitudes and as a group underestimated the actual temporal magnitudes of geological events. Catley and Novick (2009) also found that the previous biology coursework did not help students make absolute temporal-magnitude estimations or sequence the geological events according to their appearance.

Since studies on time estimation of temporal-magnitudes smaller than the human-experience range do not exist (Tretter, Jones, Andre, Negishi, & Minogue, 2006), we review studies that addressed conceptions of spatial magnitudes smaller than the human experience range. This can be meaningful due to similarities between temporal and spatial reasoning (Boroditsky, 2000; Torralbo et al., 2006). Walsh (2003) proposed a common cognitive link between space and time through magnitude; for instance, sizes of objects can be expressed and compared on the spatial scale according to magnitude, while durations of events can be expressed on a similar temporal scale according to magnitude.

In representing quantities of a particular type with a scale, numbers can be put into a continuum called the ''number line'' (Fischer, 2003). Gibbon and Church (1981) claimed that the number line is linear, and that error increases when estimating numbers far from the center of the line. However, Dehaene, Izard, Spelke, and Pica (2008) considered the number line logarithmic, meaning that the perceived distance between two adjacent numbers decreases when the numbers become larger (underestimation) and increases when the numbers become smaller (overestimation). Longo and Lourenco (2007) called this phenomenon the *compression effect* towards the human-experience range. This compression effect is found in underestimations of durations for large temporal magnitudes (Catley & Novick, 2009) as well as objects' sizes below and above human experience range (Price & Lee, 2009). In asking subjects to generate objects with various spatial magnitude categories, Tretter et al. (2006b) found a decrease in accuracy outside of the everyday range of 1 mm to 100 m for all age groups. In addition, teachers (Jones, Tretter, Taylor, & Oppewal, 2008) and students (Jones et al., 2007) alike were less accurate with small spatial scale items than large spatial scale items. Jones and Taylor (2009) identified that the scientists developed the sense of scale through in-school experiences such as measurement, creating models, and drawing maps as well as out-of-school experiences associated with physical movement.

One can have direct knowledge of temporal magnitudes of certain durations such as the common knowledge that a day is about 24 hours. Without direct knowledge of a particular duration, one might use magnitude estimation strategies known as bootstrapping or benchmarking. Bootstrapping refers to the process of increasing the range of a scale by using the total magnitude of one scale to link to yet another scale. When performed multiple times, it creates a ladder from one scale to another. This bootstrapping strategy was mentioned in *Benchmark for Science Literacy,* which suggested students should be able to

''bootstrap their comprehension of magnitude only by a few factors of 10 at a time, perhaps grasping each new level only in terms of the previous one'' (AAAS, 1993, p. 279). Another technique of moving between different scales is to use benchmarks related to established quantities within a scale. Once the magnitude of a benchmark is known, it acts as a reference point for other nearby magnitudes. This establishes independent scales surrounding each benchmark. Experts with solid comprehension of scale make use of both bootstrapping and benchmarking strategies (Jones & Taylor, 2009). Lamon (1994) calls one such combined strategy as ''unitizing,'' which involves taking a collection of objects at one scale, grouping them together and using them to describe an object at a different scale (example: unitizing the distance light travels in a year as light year and describing interstellar distance in light years). Current literature on conceptions of magnitude is unclear on whether knowledge of the exact temporal or spatial magnitudes is needed, or whether estimation strategies are sufficient in order to recognize temporal or spatial magnitudes accurately.

Methods

Temporal-Magnitude Instrument Design

For this study, we designed a paper and pencil instrument. To sample scientific changes across temporal magnitudes, we selected 30 items from the National Science Education Standards (NRC, 1996). Some durations used in the instrument were directly mentioned in the standards, for example, ''The sun, the earth, and the rest of the solar system formed from a nebular cloud of dust and gas 4.6 billion years ago'' (NRC, p. 189).

Duration is measured or estimated by calculating the difference between two time points, $\Delta t = t_{\text{end}} - t_{\text{start}}$. For *trends*, the start time is when the process is initiated and the end time is when the process is completed. For *cycles*, duration can be represented as a period. Durations of *evolutionary changes* can be estimated as the time between when a particular evolutionary event occurred (e.g., extinction of dinosaurs) and the present time. To account for irregular changes (AAAS, 1990), we added *instances* to the instrument. We define an *instance* as a step or situation viewed as an illustrative part of collective processes. For example, ''chemical reactions can take place in time periods ranging from the few femtoseconds (10^{-15} seconds) required for an atom to move a fraction of a chemical bond distance'' (NRC, 1996, p. 179). Other *instances* included changes that occur as instantaneous responses to external stimuli, such as the blink of an eye, neural responses, and lightning. We placed durations related to kinetic movements into the *instances* category, such as the time for light to travel 1 m. A commonly taught change that was not included in this instrument was earthquakes which have hybrid characteristics between incidences and trends.

These items (Table 1) were sampled from various science domains (14 earth science, 9 life science, and 7 physical science) to represent various change types (8 *cycles*, 8 *evolutionary changes*, 7 *trends*, and 6 *instances*). For each item, students were asked to select 1 of 13 temporal-magnitude categories. The largest category was larger than billion years and the smallest was smaller than a billionth of a second. Eight items from ATOM to NEURON represented changes smaller than a 100 milliseconds (extremely small temporal magnitudes); 11 items from SHOOT to HUMAN represented those occurring between more than a 100 milliseconds and a 100 years (previously defined as the human-experience range); 11 items from NEWTON to UNIVERSE represented those larger than a 100 years (extremely large temporal magnitudes).

We designed this instrument to test whether students' responses to the items could form a single measurement scale on the temporal-magnitude recognition construct. We use ''recognition,'' as opposed to perception or conception of durations, since students were asked to choose a temporal-magnitude category for each scientific change listed in the instrument. Even though the knowledge of a scientific change includes the knowledge of its temporal magnitude, this instrument was not designed to measure students' overall knowledge or understanding of scientific changes listed in the instrument. Rather, we conceptualized the temporal-magnitude recognition construct as the common cognition required across all 30 items.

In order to further clarify the nature of the construct, we added an answer explanation part to the instrument asking whether students answered the temporal-magnitude of each item because they ''knew the answer,'' ''estimated from other processes,'' or ''randomly guessed.'' We asked this set of questions to examine whether the temporal-magnitude recognition construct mainly relied on students' exact knowledge of the temporal-magnitude or their use of temporal-magnitude estimation strategies such as benchmarking

Table 1
Item description

Item ID	Item Description	Approximate Temporal Magnitude	Change Type	Science Domain
ATOM	Shortest interval measured by atomic clock	100 attoseconds	C	P
CHEMRC	Time for the fastest chemical reactions	A few femtoseconds	I	P
LIGHT1M	Time for light to travel 1 m	3.3 nanoseconds	I	P
RADIO	Shortest half life of a radioactive isotope	1.2 microseconds	T	P
BLINK	Eye blink	50–60 microseconds	I	B
FLASH	Lightning flash	100 microseconds	I	P
WING	Hummingbird wing flap	20 milliseconds	C	B
NEURON	Typical neuron response to stimulus	32 milliseconds	I	B
SHOOT	Appearance of a shooting star	A few seconds	T	P
BREATH	A breath	3–4 seconds	C	B
ECLIPS	Length of solar eclipse	2–10 minutes	C	ES
DIGEST	Digestion of food in human body	6–8 hours	T	B
TIDE	High and low tidal cycle	12 hours	C	ES
SPOIL	Time for pasteurized milk to spoil in fridge	2 weeks	T	B
PHASE	Cycle length of the phases of the moon	29.5 days	C	ES
SEASON	A season	3 months	C	ES
EMBRYO	Incubation period of human embryo	9 months	T	B
SUNSPOT	Sun spot cycle	11 years	C	ES
HUMAN	Human lifespan	70 years	T	B
NEWTON	Time since Newton's laws were formalized	300 years	—	P
TREE	Age of oldest known tree	9,550 years	T	B
SAPIENS	Time since the appearance of modern humans	195,000 years	E	ES
GALAXY	Time to fly to the nearest galaxy at light speed	2.56 million years	I	ES
MOUNT	Mountain formation	13 million years	E	ES
DINO	Time since the extinction of dinosaurs	65 million years	E	ES
FUEL	Formation of fossil fuel	300 million years	E	ES
LIFE	Time since the appearance of first life on earth	3.5 billion years	E	ES
EARTH	Age of Earth	4.5 billion years	E	ES
SUN	Age of Sun	4.5 billion years	E	ES
UNIVERSE	Age of Universe	13.8 billion years	E	ES

Note. T, trend; C, cycle; E, evolutionary change; I, incidence; P, physical science; B, biological science; ES, earth and space science.

and bootstrapping. The temporal-magnitude instrument with the answer explanation part is available as supplementary material accompanying the online article.

Data Collection

Data were collected in two phases. In the first phase, we recruited 12 instructors from 10 higher education institutions in the United States. According to the Carnegie Classification, the 10 institutions consisted of 3 Tier-1 universities with doctoral programs, 3 Liberal Arts colleges (1 Tier-1, 1 Tier-3, and 1 Tier-4), 2 universities with Masters' programs (1 Tier-1 and 1 Tier-3), 1 community college, and 1 Tier-1 specialty college with engineering programs. At the time of data collection, these instructors taught astronomy-related courses, except one who taught bio-molecular science and science and society. They administered the temporal-magnitude instrument to their students ($n = 419$) towards the end of their course. These courses did not address temporal magnitudes of scientific changes as a main course objective.

In the following semester, 95 students in a cosmic evolution course took the temporal-magnitude instrument with the answer explanation part on the first day of the course. The cosmic evolution course was offered at a Tier-1 University with doctoral programs in the Northeastern part of the United States. On the last day of the course, 59 students took the same instrument with the answer explanation part. The number of the students who took the instrument as a posttest was smaller than the number that took it as a pre-test because the posttest was administered on the last day of the course where student attendance was not required.

A total of 514 students participated. These students consisted of 56.4% male and 43.6% female; the represented majors included 43.0% social science, 24.6% science and engineering, 20.8% language and humanities, and 11.7% undecided majors. According to self-report, 80.3% took a course on physics, 94.5% on chemistry, and 96.9% on biology during high school. Students took about 10–15 minutes to complete the temporal-magnitude instrument.

The cosmic evolution course was taught by professors from five different science disciplines: astronomy, geology, chemistry, biology, and anthropology. The course detailed the scientific account of the origin and evolution of organized structures found in the universe. An astrophysicist taught the period of time from the Big Bang to the formation of the solar system; a geologist described the formation of earth and current climate system; a chemist described the evidence and mechanisms related to how life began; a biologist described how single cell organisms evolved to become multi-cellular; an anthropologist detailed how *homo sapiens* evolved from ancient ancestors. Therefore, students in the course were expected to learn evolutionary changes with extremely long temporal magnitudes (most of the items representing larger than 1 million years) as well as the importance of Sun as an energy source, chemical changes associated with emergence of building blocks of life, and the role of DNA molecules in evolution. Students in the cosmic evolution course took two mid-term and a final exams. The exams were created by the course instructors and consisted of multiple-choice, short-answer, and open-ended items. The exams did not directly ask temporal magnitudes of scientific phenomena addressed in the course.

Data Analysis

For each item, we assigned a ''1'' when students correctly recognized the temporal magnitude and a ''0'' when they did not. On average, 3.9% of the answers (about 20 answers from 514 eligible answers on each item) were missing across 30 items. The largest percentage of missing answers was found on RADIO (7.2%) and the smallest was found on TIDE (1.8%). These missing answers were interpreted as not recognizing correct temporal magnitudes and thus assigned a ''0.''

To examine students' tendency to overestimate or underestimate temporal-magnitude categories from their actual magnitude categories, we computed an error variable defined as the difference between the chosen and the correct temporal categories ($\text{category}_{\text{chosen}} - \text{category}_{\text{correct}}$) for each item. For example, if a student selected the fourth temporal category for an item when the correct category was the fifth, then the value on the error variable for the student on that item was -1. Negative values on the error variable represented underestimation from actual temporal magnitudes while positive values represented overestimation.

We applied a Rasch analysis (1960/1980) on student responses to the instrument using *ConQuest* (Wu, Adams, Wilson, & Haldane, 2007). The Rasch model used in this study can be formulated as

$$P = (X_i = 1|\theta_n, \delta_i)\frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}}$$

where P is the probability of correctly recognizing the temporal magnitude of an item i given student n's ability, θ_n, and the item difficulty, δ_i. We used item difficulty estimates on the log-odds unit (logit) scale to examine in which items students had difficulty recognizing temporal magnitudes. We also used fit statistics to investigate whether the temporal-magnitude recognition ability could be treated as a single construct. Using correlations, we investigated how knowledge, estimation, and random guesses were related to item difficulty to clarify the type of cognition involved in the temporal-magnitude recognition construct measured with the instrument.

To examine instructional sensitivity of the instrument, we applied the paired samples t-test at $\alpha = 0.05$ to Rasch ability estimates of students who took the temporal-magnitude instrument before and after the cosmic evolution course. Instructional sensitivity in this study is defined as ''the tendency for an item to vary in difficulty as a function of instruction'' (Haladyna & Roid, 1981, p. 40). If the instrument was sensitive to the cosmic evolution course that targeted scientific changes with extremely long durations, students would improve their temporal magnitude recognition ability after the cosmic evolution course by mainly improving their performances on the extremely long duration items. Further, students would improve on the extremely long duration items after the course by more frequently employing the type of cognition involved than they

did before the course. We therefore compared changes in both students' temporal-magnitude selections and explanations of their selections using McNemar tests at $\alpha = 0.05$.

We collected all of the student exam scores and calculated average exam scores as an indicator for their overall knowledge about scientific phenomena addressed during the cosmic evolution course. We calculated correlation coefficients to investigate how students' knowledge of the scientific phenomena was associated with their temporal-magnitude recognition ability measured after the course as well as their temporal-magnitude recognition ability difference before and after the course.

Results

Student Temporal-Magnitude Recognition Patterns

Students correctly categorized an average of 13.9 items ($SD = 4.5$) out of 30 scientific changes, with scores ranging from 0 to 24. Figure 1 shows the percentages of students who chose correct categories across the 30 items. The most accurately recognized item was the human life-span (HUMAN, 85.6% correct) while the least accurate was the typical neuron response time to external stimuli (NEURON, 11.7% correct). Correct percentages were higher overall within the human-experience range and gradually diminished towards both ends with a slight increase at the small temporal-magnitude end and a large increase at the large temporal-magnitude end. These overall patterns are largely consistent with spatial scale studies (Jones et al., 2007, 2008; Tretter, Jones, & Minogue, 2006).

While correct percentages of all eight items below the human-experience range were rather consistent, those across the rest of temporal categories showed large variations. Apparently, some durations in the human-experience range such as SUNSPOT were more difficult than other durations in the longer than human-experience range (SAPIENS, MOUNT, and DINO). This indicates that the temporal magnitude of scientific change alone cannot predict whether or not students would successfully recognize the temporal magnitude.

Figure 2 shows the average errors made in selecting temporal-magnitude categories. Overall, students' errors (see the length of the bars) were smaller from ECLIPSE to HUMAN in the human experience range than below and above the human experience range. Students tended to overestimate durations of smaller than 1 millisecond (from ATOM to FLASH) whereas they tended to underestimate durations for most of the other temporal-magnitude categories. This consistency in the direction of errors shows the compression effect at

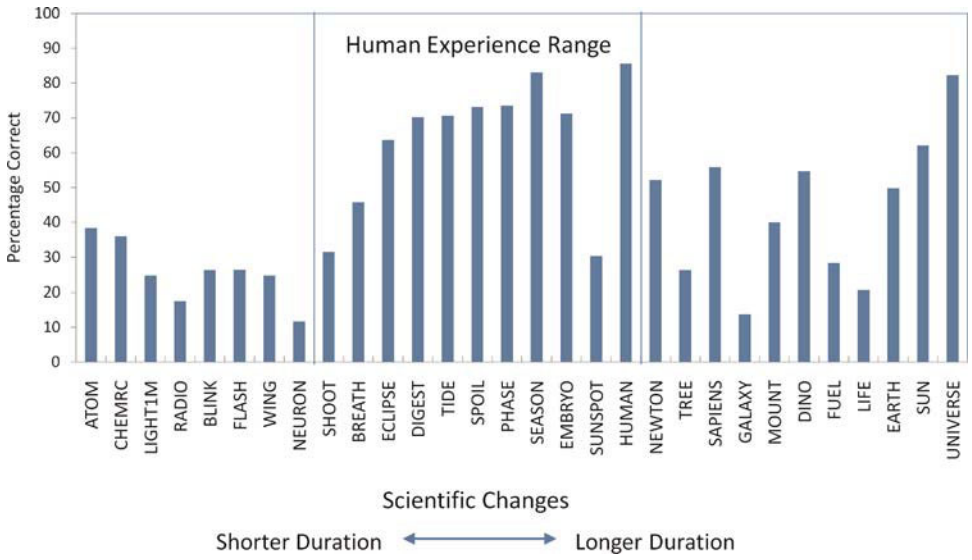

Figure 1. Correct temporal-magnitude categorization percentages across scientific changes.

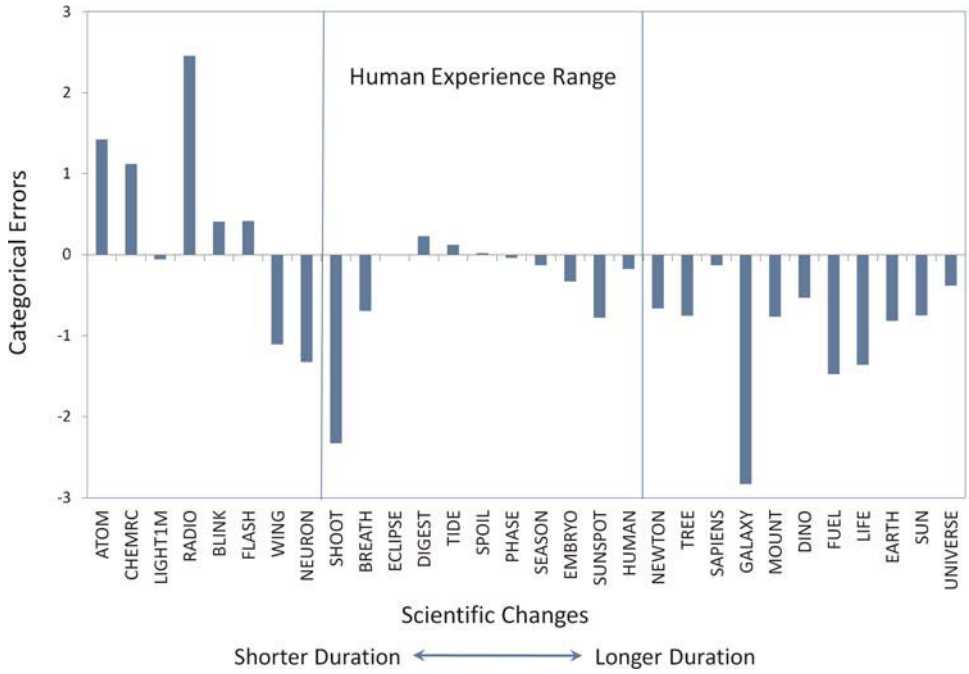

Figure 2. Temporal-magnitude categorization errors.

both ends of the temporal magnitudes. While these overall results are consistent with the psychology literature (Eisler, 1976; Ward, 1975; Zakay & Block, 1997), our results did not agree on where underestimation begins to occur. According to Eisler (1976), it should begin to occur at SHOOT (appearance of a shooting star), which has a duration of a few seconds. In this study, students started to change from overestimation to underestimation beginning at WING, with a duration of about 20 milliseconds. According to Ward (1975), overestimation occurs with durations at or shorter than 100 milliseconds.

Establishing a Measurement Scale Based on Rasch Analysis

Item Characteristic Curve (ICC). The ICC of an item shows the probability of a correct response given student ability on the construct. Since student ability varies on the construct, the probability is drawn as a curve over the low to high ability range. Figure 3 shows the ICC for the LIFE item (time since the appearance of the first life form on Earth). The vertical axis represents the probability of a correct response to the LIFE item, and the horizontal axis represents student ability ranging from −2.0 to 2.0. The larger the value on the horizontal axis, the more able the student on the temporal-magnitude recognition construct. The solid line in Figure 3 stands for the expected probability of a correct response given student ability, while the dotted line stands for the empirical probability calculated from the actual student responses. For the least able students at the logit value of −1.0, their probability of answering correctly on this item is about 0.1. As students are more able, their probability of a correct response to the LIFE item increases monotonically. For students at the logit 1.0 level, the expected probability of a correct response to the LIFE item increases to 0.45.

Item Fit Statistics. Using the marginal maximum likelihood estimation method, *ConQuest* computes two types of fit statistics called *outfit* and *infit* to indicate the discrepancies between the modeled probability (the solid line on Figure 3) and the actual data (the dotted line on Figure 3). How well actual data fit the expected probability can be represented with fit statistics, either as chi-square statistics divided by degrees of freedom (mean square) or as standardized *t*-statistics (Wilson, 2005). Here, we use infit/outfit mean square values to describe how well the items fit the expected probabilities based on the Rasch model. Infit detects

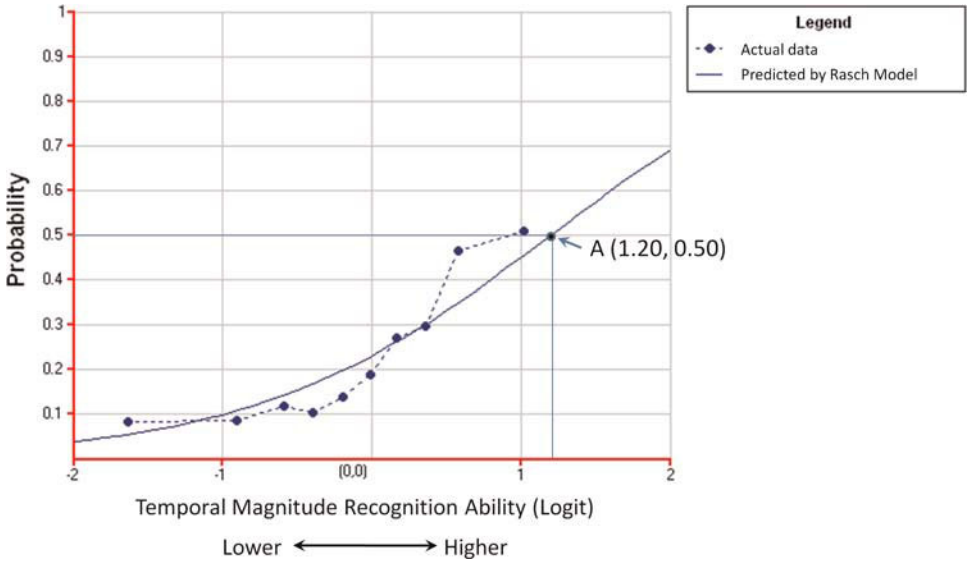

Figure 3. Item characteristic curve: LIFE.

unexpected patterns of responses to items while outfit detects responses to items that are far beyond or below a student's ability. To indicate the model fit, the expected value for infit/outfit mean square is +1.0 and the infit/outfit statistics can range from 0.0 to positive infinity. The acceptable value range for the infit/outfit mean square for data-model fit is between 0.7 and 1.3 (Bond & Fox, 2007). The LIFE item shown in Figure 3 has an outfit mean square value of 0.94 and an infit mean square value of 0.97. Table 2 shows both infit and outfit mean square values for all items. The outfit mean square values ranged from 0.75 (SEASON) to 1.24 (LIGHT1M). The infit mean square values ranged from 0.86 (SEASON) to 1.16 (CHEMRC). Based on the 0.7–1.3 acceptable range, all 30 items in the instrument item could be used to predict student ability on a single construct scale according to the Rasch model.

Item Difficulty. The item difficulty estimate of an item is determined by the intersection point between the ICC of the item and the 0.50 probability line. For instance, the ICC of the LIFE item intersects with the 0.50 probability line at (1.20, 0.50). See the Point A in Figure 3. The horizontal axis value of 1.20 is the item difficulty estimate for the LIFE item. Table 2 lists item difficulty values of all 30 items. Item difficulty values ranged from −2.22 (HUMAN) to 2.09 (NEURON). Since each item was assigned a unique estimate of item difficulty, we used ANOVA to compare mean item difficulties of three item groups defined by temporal-magnitude. The mean item difficulty estimate for correctly recognizing durations smaller than the human-experience range was 1.07 ($n = 8$, $SD = 0.57$), while the mean for correctly recognizing durations at the human-experience range was −0.87 ($n = 11$, $SD = 0.97$). The mean item difficulty for changes larger than the human-experience range was 0.10 ($n = 11$, $SD = 1.07$). ANOVA indicates that these mean item difficulty values were significantly different, $F(2, 27) = 10.24$, $p < 0.001$. Post hoc tests indicate that the mean item difficulty of durations smaller than the human-experience range was significantly lower than that of durations at the human-experience range ($p < 0.001$) and was not significantly different from that of durations larger than the human-experience range ($p = 0.053$). Tukey's HSD tests also indicate no significant mean difference between durations larger than and at the human experience range ($p = 0.078$).

Measurement Scale for the Temporal-Magnitude Recognition Construct on Wright Map. Figure 4 shows a Wright Map that represents the distribution of students according to their ability values and that of items according to their item difficulty values within a range of −3.0 (very low ability/very easy) to 3.0 (very high ability/very difficult). Students placed higher on the measurement scale have higher temporal scale

Table 2
Rasch analysis results with percent frequencies on knowledge, estimation, and random guessing across 30 scientific changes

	Rasch Analysis Results ($N = 514$)				Answer Explanation (%, $n = 95$)		
Item ID	Item Difficulty	Error	Outfit Mean Square	Infit Mean Square	Knew	Estimated	Guessed
ATOM	0.35	0.071	1.05	1.06	7.8	33.3	58.9
CHEMRC	0.39	0.072	1.22	1.16	7.8	33.3	58.9
LIGHT1M	1.14	0.077	1.24	1.06	14.4	55.6	30.0
RADIO	1.55	0.082	1.10	1.05	2.2	29.7	68.1
BLINK	0.97	0.075	1.19	1.10	22.8	58.7	18.5
FLASH	1.00	0.076	1.15	1.08	12.2	62.2	25.6
WING	1.07	0.076	1.13	1.04	8.8	53.3	37.8
NEURON	2.09	0.089	1.12	1.01	8.0	45.5	46.5
SHOOT	0.68	0.073	1.06	1.06	9.0	44.9	46.1
BREATH	0.01	0.071	0.93	0.96	45.1	39.6	15.4
ECLIPS	−0.79	0.072	0.91	0.93	25.3	52.7	22.0
DIGEST	−1.13	0.075	1.02	1.02	48.9	41.3	9.8
TIDE	−1.14	0.075	1.04	1.03	47.8	40.2	12.0
SPOIL	−1.28	0.076	1.04	1.00	34.4	44.4	21.1
PHASE	−1.32	0.076	0.94	0.96	48.3	33.7	18.0
SEASON	−1.95	0.083	0.75	0.86	70.8	22.5	6.7
EMBRYO	−1.20	0.075	0.90	0.91	52.2	26.1	21.7
SUNSPOT	0.76	0.074	1.02	1.02	4.4	26.4	69.2
HUMAN	−2.22	0.086	0.84	0.90	76.1	18.5	5.4
NEWTON	−0.32	0.071	0.91	0.93	46.1	40.4	13.5
TREE	0.98	0.076	1.09	1.06	17.4	50.0	32.6
SAPIENS	−0.42	0.071	1.01	1.01	19.6	44.6	35.9
GALAXY	1.86	0.086	1.21	1.04	6.7	50.0	43.3
MOUNT	0.28	0.071	1.00	1.01	8.8	52.7	38.5
DINO	−0.40	0.071	0.89	0.91	35.6	43.3	21.1
FUEL	0.82	0.074	0.95	0.96	8.8	45.1	46.2
LIFE	1.20	0.078	0.94	0.97	25.3	48.4	26.4
EARTH	−0.21	0.071	0.93	0.94	42.9	44.0	13.2
SUN	−0.73	0.072	0.90	0.92	32.2	40.0	27.8
UNIVERSE	−1.99	0.408	0.93	0.97	47.3	38.5	14.3

recognition abilities. Items placed higher are more difficult for students to choose a correct temporal category than those below them. On this measurement scale, the mean of students' temporal-scale recognition ability was −0.19 with a standard deviation of 0.63. The minimum ability estimate was −2.13 and the maximum ability estimate was 1.21. The most difficult item was the NEURON while the easiest was HUMAN.

In the Rasch model, the probability of a student's correct response to an item depends on the student's ability and the item's difficulty. If a student's ability estimate is higher than the difficulty estimate of an item, the student has a larger than 50% chance of answering that item correctly. On the other hand, if a student's ability estimate is lower than the item difficulty estimate, the student has a less than 50% chance of answering that item correctly. That is, students with an ability estimate of 1.20 have a 50% chance of choosing the correct temporal-magnitude category for the LIFE item (item difficulty = 1.20), a less than 50% chance for the RADIO item (item difficulty = 1.55), and a more than 50% chance for the DINO item (item difficulty = −0.40).

The Rasch analysis produces two types of reliability indicators for the measurement scale: one for items and the other for persons. For the scale on the temporal-magnitude recognition construct, the person separation reliability was 0.77 while the item separation reliability was 0.99. The person separation reliability was lower than the item separation reliability since the former was based on 30 items while the latter was based on 514 students. The person separation reliability is analogous to the traditional Cronbach's alpha value.

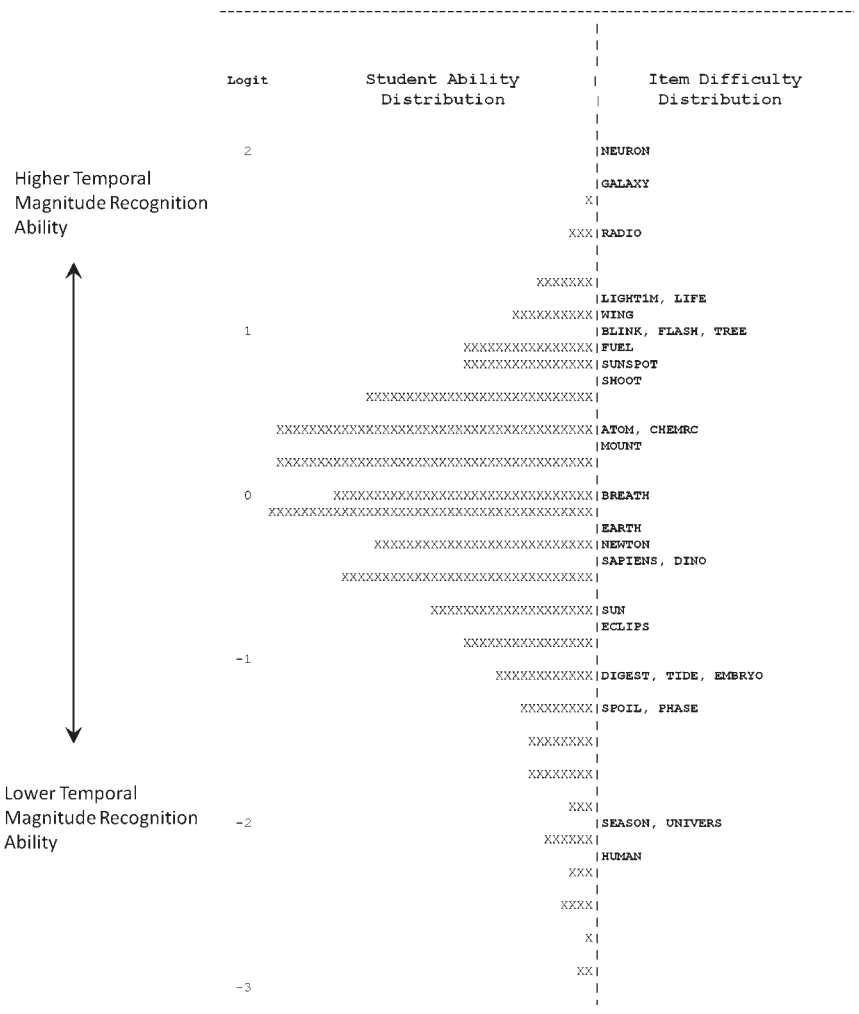

Figure 4. Wright Map for the temporal-magnitude recognition measurement scale.

Knowledge, Estimation, and Random Guessing

We analyzed data from the 95 students who responded to the answer explanation part before the cosmic evolution course. There was no significant difference in the temporal-magnitude recognition ability estimates between those who responded to the answer explanation part ($n = 95$, $M = -0.31$, $SD = 0.56$) and those who did not ($n = 419$, $M = -0.27$, $SD = 0.62$), $t(512) = 0.58$, $p = 0.56$. Based on this information, we assumed that the students who answered the answer explanation part could approximate the entire sample's responses. According to Table 2, the item with the largest percentage of students who knew the temporal-magnitude was HUMAN (76.1%) and the item with the smallest percentage was RADIO (2.2%). Table 2 shows that 25% or fewer students knew the answers for half of the 30 items, including all eight items shorter than the human experience range.

We calculated Pearson correlations between item difficulty and each of the percent frequencies for knowledge, estimation, and guess categories. According to Cohen's guide on interpreting the strength of correlations (Cohen, 1988), knowledge of temporal-magnitude was strongly negatively correlated with item

difficulty ($r = -0.86$, $p < 0.001$), meaning the larger the percentage of students who knew the item's temporal magnitude the easier the item. On the other hand, random guessing was strongly positively correlated with item difficulty ($r = 0.68$, $p < 0.001$). So was estimation ($r = 0.52$, $p < 0.01$). This means that students more frequently estimated or randomly guessed on more difficult items. It is surprising that estimating the temporal magnitude of a scientific change based on that of other known scientific changes was correlated positively with item difficulty. Estimation strategies such as benchmarking and bootstrapping might not be enough for students to recognize temporal magnitudes accurately.

Instructional Sensitivity

We compared students who took the temporal-magnitude instrument both before and after the cosmic evolution course ($n = 59$). The mean ability estimate measured before the cosmic evolution course was -0.22 ($SD = 0.63$) while the mean on the posttest was 0.07 ($SD = 0.49$). This improvement was significant using a paired t-test, $t(58) = 3.17$, $p < 0.01$. The effect size was 0.52, a medium impact (Cohen, 1988). Table 3 lists student performance changes in all 30 items. Before and after the cosmic evolution instruction, the students selected different temporal magnitude categories as high as 50.8% on the NEWTON item and as low as 3.4% on the UNIVERS item. On 25 out of 30 items, students who changed from incorrect to correct

Table 3
Changes in students' performances and explanations before and after the cosmic evolution course (n = 59)

	Performances on Temporal-Magnitude Recognition			Explanations for Temporal-Magnitude Selection		
Item ID	Incorrect → Correct	Correct → Incorrect	McNemar Test, p	Estimated/ Guessed → Knew	Knew → Estimated/ Guessed	McNemar Test, p
ATOM	11	8	0.65	4	5	1.00
CHEMRC	15	8	0.21	8	0	**
LIGHT1M	8	8	1.00	9	7	0.80
RADIO	7	7	1.00	3	0	0.25
BLINK	13	11	0.84	11	7	0.48
FLASH	12	7	0.36	8	8	1.00
WING	10	9	1.00	6	4	0.75
NEURON	3	8	0.23	8	4	0.39
SHOOT	13	4	*	3	2	1.00
BREATH	5	9	0.42	9	10	1.00
ECLIPS	12	9	0.66	10	6	0.45
DIGEST	10	9	1.00	9	6	0.61
TIDE	6	5	1.00	8	9	1.00
SPOIL	13	11	0.84	15	7	0.13
PHASE	14	7	0.19	9	8	1.00
SEASON	7	3	0.34	12	7	0.36
EMBRYO	15	8	0.21	11	6	0.33
SUNSPOT	18	6	*	17	3	**
HUMAN	11	2	*	10	8	0.81
NEWTON	22	8	*	10	7	0.63
TREE	11	8	0.65	11	8	0.65
SAPIENS	13	12	1.00	19	2	***
GALAXY	7	12	0.36	14	3	*
MOUNT	14	4	*	13	6	0.17
DINO	12	10	0.83	14	6	0.12
FUEL	12	9	0.66	5	3	0.73
LIFE	18	3	***	22	7	**
EARTH	15	5	*	14	4	*
SUN	13	9	0.52	19	3	***
UNIVERSE	2	0	0.50	18	2	***

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

answers outnumbered those who changed from correct to incorrect answers. According to McNemar tests, seven of these changes including four extremely long durations were statistically significant. The other significant positive changes on SHOOT (appearance of a shooting star), SUNSPOT (sunspot cycle), and HUMAN (human lifespan) can also be related to the course content: meteorites as a main source of evidence for studying the conditions of the early Universe or the early Solar System, the Sun as a main energy source driving evolutionary changes on the Earth, and the origin and evolution of the human race.

Similarly, we analyzed how students' self-reports on the answer explanation part changed before and after the course. If students improved on the temporal-magnitude recognition construct, then it should coincide with an increase in their self-reported knowledge of the temporal magnitudes, because only knowledge was significantly negatively correlated with item difficulty. Therefore, we expected that a larger number of students would change from estimation/random guessing to knowledge than from knowledge to estimation/random guessing on the items addressed in the cosmic evolution course. Table 3 shows that significant changes occurred in eight items, including six extremely long duration items. As the cosmic evolution focused on the history of the Universe from the Big Bang to the present time, these significant changes in students' self-report were well aligned with the course content. In addition, significant change in the CHEMRC item might be related to the fact that the course spent time on the formation of organic molecules (building blocks of life) in the primitive atmosphere of the Earth. The significant change on the Sunspot cycle could be related to the importance of the Sun as a driving force for the Earth history.

Knowledge of Scientific Phenomena

The knowledge of the scientific phenomena was weakly and positively but not significantly correlated with the temporal magnitude recognition ability measured after the course, $r = 0.26$, $p = 0.06$, and the difference between before and after the course, $r = 0.16$, $p = 0.25$.

Discussion

It is suggested that ''change should not be taught as a separate subject. At every opportunity throughout the school years, the theme of change should be brought up in the context of the science, mathematics, or technology being studied'' (AAAS, 1993, p. 272). If this suggestion were well implemented, recognition of temporal magnitudes associated with durations of scientific changes would be expected as a cumulative learning outcome over multiple science courses throughout formal school years. We developed an instrument to measure this temporal-magnitude recognition ability. We tested the instrument with college students using scientific changes that should have been addressed through K-12 science education. To characterize the temporal-magnitude construct using the instrument, we collected validity evidence from psychological, psychometrical, and instructional sources.

Theories and previous empirical evidence on psychological perceptions of durations within the range of milliseconds to a few days can be extended to all temporal magnitudes of the items used in this study. College students' overall temporal-magnitude recognition ability is not equally well developed across all temporal magnitudes. The temporal-magnitude recognition ability peaks in the human-experience range and becomes weaker over temporal magnitudes above and below that range. This overall profile is consistent with how subjects perceive spatial magnitudes (Jones et al., 2007, 2008; Tretter, Jones, & Minogue, 2006) as well as geological events (Trend, 2001).

As predicted in the logarithmic theory of the number line (Dehaene et al., 2008), the way students recognize temporal magnitudes appears to be logarithmic, rather than linear. The logarithmic theory of the mental number line predicts a rather accurate middle range before the range of underestimation or overestimation appears. Our analysis shows that errors are smaller in the human experience range than the extremely short and the extremely long ranges. Students tend to overestimate extremely short durations and underestimate extremely long durations. Catley and Novick (2009) found the underestimation of extremely long durations when college students generated temporal magnitudes associated with extremely long durations in an open-ended item format and called this underestimation forward-telescoping. It appears that underestimation of extremely large temporal magnitudes is consistently found, regardless of how student responses are elicited—that is, choosing or generating a temporal magnitude associated with scientific

change. Furthermore, our results show that the distortion also occurs on the other side of temporal magnitude in the form of overestimation. Combining distortions on both ends of the temporal-magnitude continuum, students' temporal-magnitude recognition ability shows the compression effect (Longo & Lourenco, 2007).

Since students' treatment of temporal magnitude appears to be logarithmic, we caution the use of visualizations that alter the actual temporal magnitudes by speeding up or slowing down the durations of scientific changes. Illustrating the entirety of changes that cannot be perceived by humans is important, but it should be done so with accurate information on the actual temporal magnitudes. Otherwise, off-scale viewing of scientific changes may further reinforce the compression effect students already hold, potentially impeding student learning of both extremely short and long changes.

A Rasch analysis is used to investigate how student responses to the items in the instrument contribute to the measurement of the underlying temporal-magnitude recognition ability. Rasch analysis results indicate that all items in the instrument can be used to form a single measurement scale. On this measurement scale, the temporal-magnitude recognition ability of the entire student sample ranges from −2.13 to 1.21 out of the −3.0 to 3.0 scale. This means that the entire sample we used does not include students with very high temporal-magnitude recognition abilities. Despite the call for students' competency in the use of powers of 10 to grasp temporal magnitudes of scientific changes by the 12th grade (AAAS, 1993), current college students do not appear to have an accurate understanding of temporal magnitudes beyond the human experience range. This can be expected because school science curricula are typically organized to address scientific concepts and topics in similar temporal magnitudes and rarely provide opportunities for students to compare scientific changes at various temporal magnitudes.

Some science disciplines such as physics, cosmology, and geology may be in a better position to address differences in temporal magnitudes as compared to other disciplines such as chemistry or biology. Strengthening the temporal aspect of scientific changes across various science courses is needed. Another approach is to use a ''short capstone course on the subject of change . . . after they [students] have a storehouse of experience with change of different kinds'' (AAAS, 1993, p. 272). The cosmic evolution course is such a course where college students learn about the history of the universe to the present while encountering various magnitudes of scientific change. The pre-posttest comparison results indicate that students can enhance their temporal-magnitude recognition ability through increasing knowledge of the temporal magnitude.

More importantly, Rasch analysis results indicate that extremely short durations are more difficult for students to recognize temporal magnitudes accurately. Coincidentally, most students self-reported as not having knowledge on these small duration items. Considering the current status of science education relating to students' understanding of subatomic phenomena (Castellini et al., 2007; Roco, 2003; Tretter, Jones, & Minogue, 2006), finding easy items for extremely small durations will be a difficult task.

Temporal-magnitude recognition not only depends on students' perception of durations and magnitudes, but also on their knowledge of scientific phenomena. For example, though fewer errors are observed across the items in the human-experience range as compared to those in the other two extreme ranges, there are a few human experience range items where students have difficulty identifying correct temporal magnitudes, such as the sunspot cycle. Likewise, there are items that are not in the human experience range but for which students are able to identify correct temporal magnitudes, such as extinction of dinosaurs or age of the universe. However, it is unclear as to how much and what type of knowledge about scientific phenomena is needed for students to have a solid temporal-magnitude recognition ability. In this study, students' knowledge of scientific phenomena measured with the exams administered by the cosmic evolution course instructors was not significantly correlated with their temporal-magnitude recognition ability measured after the course. This indicates that the temporal-magnitude recognition construct might potentially tap on a different aspect of cognition from the knowledge of the related scientific phenomena itself.

Significant, strong, negative correlations between knowledge of temporal magnitude and item difficulty confirm the importance of having direct knowledge of temporal magnitude. On the other hand, estimating temporal magnitudes using other known changes is positively correlated with the item difficulty. Surprisingly, the correlation of estimating with item difficulty is similar to that of random guessing. These findings along with the logarithmic nature of magnitude perception suggest that bootstrapping and benchmarking may not work effectively without information on exact temporal magnitudes.

The findings in this study are limited to the items used in this study; electing other sets of items may produce different results because students' knowledge with the items plays a role in selecting correct temporal-magnitude categories. We did not test other types of instruments on temporal magnitudes, such as generating temporal magnitudes to a given set of scientific changes or generating scientific changes to a given set of temporal magnitudes in an open-ended format. Therefore, this study does not answer whether the instrument we used works better than the other types of instruments on temporal magnitudes. We encourage other researchers to validate other types of instruments using Rasch analysis for comparison. Though the study subjects represent a broad spectrum of students in terms of gender, major, and college setting, they are not randomly drawn from the general population. In this study, instructional sensitivity of the instrument was tested with the cosmic evolution course that targeted extremely long durations. An additional study is needed to test its sensitivity to instruction targeting other temporal magnitude ranges such as nano-scale science. A paper-and-pencil test is used to identify overall patterns related to temporal-magnitude recognition; qualitative studies on a smaller set of students using interviews and visual stimuli can shed light on how and why students recognize temporal magnitudes of certain durations in their own accounts.

Conclusion

Establishing the order of events and estimating the duration of the events are important elements of science, as well as everyday life. The temporal order of some scientific phenomena has been taught in geology, cosmology, and biological evolution. Time dependence of some science topics has been used to structure curricular content in terms of constancy and change. Since time is involved in every scientific phenomenon imaginable, time-related concepts such as order, duration, and time dependence has a potential for even broader integration with current K-12 science topics and disciplines. Results of this study indicate that (1) the temporal-magnitude recognition ability is a measurable construct, (2) extremely small duration items are significantly more difficult for students to recognize accurate temporal magnitudes than other duration range items, (3) knowledge, not estimation, of the magnitude contributes to the measurement of the construct with the instrument used, (4) the instrument is sensitive to instruction designed to improve the construct, and (5) the temporal-magnitude recognition ability is not significantly correlated with knowledge about the related scientific changes.

Further research is needed to disentangle what relationships students' temporal magnitude recognition ability has with other types of cognition such as scientific knowledge, number sense, and psychological time perceptions. Research is also needed to investigate whether and how the temporal-magnitude recognition ability can help or hinder students learn science. A positive relationship is plausible, as the ability to distinguish scientific changes across temporal magnitudes can reduce misconceptions related to assigning everyday scale properties to extreme ranges. Once the positive relationship is confirmed, it is important to design interventions that help students improve their temporal-magnitude recognition ability. Considering the increased interests in teaching nanoscale science (Roco, 2003), students' natural interests in the origins of the Universe and the life (NRC, 1996), and students' already developed abilities with qualitative temporal reasoning by ages 10 or 11 (Piaget, 1969), it is reasonable to promote time as a major unifying theme in science learning (Montangero, 1996). Taking into account a strong connection between temporal and spatial reasoning, our findings on students' recognition of temporal magnitudes along with findings on students' conception of spatial magnitudes (Tretter, Jones, & Minogue, 2006) can jointly support Walsh (2003)'s hypothesis on a common link between spatial and temporal reasoning through quantity.

This material is based upon work supported by the Bernstein Faculty Fellows program at Tufts University. The cosmic evolution course studied in this article was supported by the Howard Hughes Medical Institute (HHMI) to Dr. David Walt at Tufts University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Bernstein Faculty Fellows program or the HHMI. The authors gratefully acknowledge Dr. Eric Chaisson who provided scientific insights on the topics of time and change as well as the college science course instructors who provided access to their classrooms for data collection. The authors also thank Meredith Knight, Polly Donovan, Kristen B. Wendell, and Jenny Konon for their contributions to various phases of this research.

References

American Association for the Advancement of Science [AAAS]. (1990). Science for all Americans. New York: Oxford University Press.

American Association for the Advancement of Science [AAAS]. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Barnett, J.E. (1999). Time's pendulum: From sundials to atomic clocks, the fascinating history of timekeeping and how our discoveries changed the world. San Diego, CA: Harcourt Brace & Company.

Block, R.A. (1979). Time and consciousness. In: G. Underwood & R. Stevens (Eds.), Aspects of consciousness: Vol. I. Psychological issues. (pp. 179–217). London: Academic Press.

Block, R.A. (1990). Models of psychological time. In: R.A. Block (Ed.), Cognitive models of psychological time. (pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bracewell, R.N. (1999). The Fourier transform & its applications. New York: McGraw-Hill.

Bond, T.G., & Fox, C.M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75, 1–28.

Brown, S.W. (1985). Time perception and attention: The effects of prospective versus retrospective paradigms and task demands on perceived duration. Perception & Psychophysics, 38, 115–124.

Campbell, N.R., (1920) Physics the elements: Cambridge University Press.

Castellini, O.M., Walejko, G.K., Holladay, C.E., Theim, T.J., Zenner, G.M., & Crone, W.C. (2007). Nanotechnology and the public: Effectively communicating nanoscale science and engineering concepts. Journal of Nanoparticle Research, 9, 183–189.

Catley, K.M., & Novick, L.R. (2009). Digging Deep: Exploring college students' knowledge of macroevolutionary time. Journal of Research in Science Teaching, 46(3), 311–332.

Chaisson, E.J. (2001). Cosmic evolution: The rise of complexity in Nature. Cambridge, MA: Harvard University Press.

Chaisson, E.J. (2006). Epic of evolution: Seven ages of the cosmos. New York: Columbia University Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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, 1217–1220.

Dodick, J., & Orion, N. (2003a). Geology as an historical science: Its perception within science and the education system. Science & Education, 12, 197–211.

Dodick, J., & Orion, N. (2003b). Cognitive factors affecting student understanding of geologic time. Journal of Research in Science Teaching, 40(4), 415–442.

Darrigol, O. (2003). Number and measure: Hermann von Helmholtz at the crossroads of mathematics, physics, and psychology. Studies in History and Philosophy of Science, 34, 515–573.

Einstein, A. (2005). Relativity: The special & the general theory. New York: Pearson Education, Inc.

Eisler, H. (1976). Experiments on subjective duration 1868–1975: A collection of power function exponents. Psychological Bulletin, 83, 1154–1171.

Fischer, M.H. (2003). Spatial representations in number processing-evidence from a pointing task. Visual Cognition, 10(4), 493–508.

Fraisse, P. (1984). Perception and estimation of time. Annual Review of Psychology, 35, 1–36.

Gibbon, J., & Church, R.M. (1981). Time left: Linear versus logarithmic subjective time. Journal of the Experimental Analysis of Behavior, 7, 87–107.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 18(1), 39–53.

Hinshaw, G. (2009). Five-year Wilkinson microwave anisotropy probe (WMAP) observations: Data processing, sky maps, and basic results. Astrophysical Journal Supplement, 180, 225–245.

Joint Committee for Guides in Metrology. (2008). International vocabulary of metrology-basic and general concepts and associated terms. Geneva, Switzerland: International Organization for Standardization.

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., & Taylor, A. (2009). Developing a sense of scale: Looking backward. Journal of Research in Science Teaching, 46, 460–475.

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.

Lamon, S. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In: G. Hartel & J. Confrey (Eds.), The development of multiplicative reasoning in the learning of mathematics. (pp. 89–122). Albany, NY: State University of New York Press.

Longo, M.R., & Lourenco, S.F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. Neuropsychologia, 45, 1400–1406.

McPhee, J. (1981). Basin and range. New York: Farrer, Straus and Giroux.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.), Educational measurement (3rd ed.). (pp. 13–103). New York: Macmillan.

Michell, J. (1999). Measurement in psychology. New York: Cambridge University Press.

Montangero, J. (1996). Understanding changes in time: The development of diachronic thinking in 7 to 12 year old children. New York: Taylor & Francis.

National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.

Newton, I. (1995). The Principia. New York: Prometheus Books.

Piaget, J. (1969). The child's conception of time. New York: Ballantine Books.

Price, A., & Lee, H.-S. (2009). Exploring relationship between college students' categorization of spatial and temporal concepts of scale. San Diego, CA: American Educational Research Association.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Roco, M.C. (2003). Converging science and technology at the nanoscale: Opportunities for education and training. Nature Biotechnology, 7(10), 1247–1249.

Roeckelein, J.E. (2000). The concept of time in psychology: A resource book and annotated bibliography. Westport, CT: Greenwood Press.

Torralbo, A., Santiago, J., & Lupianez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30(4), 745–757.

Trend, R. (1998). An investigation into understanding of geological time among 10 and 11 years old children. International Journal of Science Education, 20, 973–988.

Trend, R. (2000). Conceptions 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. (2001). Deep time framework: A preliminary study of U.K. primary teachers' conception of geological time and perceptions of geoscience. Journal of Research in Science Teaching, 38, 191–221.

Trend, R. (2002). Chapter 13: Developing the concept of deep time. In: V.J. Mayer (Ed.), Global science literacy. (pp. 187–201). Netherlands: Kluwer Academic Publishers.

Tretter, T., Jones, 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. (2006). Accuracy of scale conceptions in science: Mental maneuverings across many orders of spatial magnitude. Journal of Research in Science Teaching, 43(10), 1061–1085.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space, and quantity. Trends in Cognitive Sciences, 7(11), 483–488.

Ward, L.M. (1975). Sequential dependence range in cross modality matches of duration to loudness. Perception & Psychophysics, 18, 217–223.

Whitelaw, I. (2007). A measure of all things: The story of man and measurement. New York: Quid Publishing.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M., Adams, R.J., Wilson, M., & Haldane, S. (2007). ACER ConQuest 2.0 [computer program]. Hawthorn, Australia: ACER.

Zakay, D., & Block, R.A. (1997). Temporal cognition. Current Directions in Psychological Science, 6(1), 12–16.

Zen, E. (2001). What is deep time and why should anyone care? Journal of Geoscience Education, 49(1), 5–9.