US-China Education Review B, November 2015, Vol. 5, No. 11, 689-700
doi:10.17265/2161-6248/2015.11.001

Initial Development of a Concept Inventory to Assess Size, Scale, and Structure in Introductory Astronomy*

Katharyn Ellen Ketter Nottis, Edwin (Ned) Ladd
Bucknell University, Lewisburg, USA

Alyssa Goodman, Patricia Udomprasert
Harvard University, Cambridge, USA

Research has shown that undergraduates have problems understanding astronomical concepts, especially size, scale, and structure. One way to evaluate understanding is to use concept inventories. Therefore, the purpose of this study was to begin the development of the Size, Scale, and Structure Concept Inventory (S3CI) to assess understanding of these concepts in introductory undergraduate astronomy courses for majors and non-majors. A secondary purpose was to determine the impact of a newly developed WorldWide Telescope (WWT) enhanced lab on parallax, part of a suite of WWT enriched labs for introductory astronomy courses currently under development. We present in this paper preliminary results from the first WWT-enhanced lab on parallax. In Fall 2013, a beta version of the S3CI was piloted in an introductory astronomy course at a small private university. An item analysis was done and estimates of internal consistency reliability were determined using the Kuder-Richardson Formula #20 (KR20). The impact of the newly developed lab was also evaluated using a sub-test of six questions from the S3CI.

Keywords: astronomy education, concept inventory, conceptual understanding, WorldWide Telescope (WWT), parallax

Introduction

Prior knowledge has a key effect on what and how much students learn (Shuell, 1992; Smith, diSessa, & Roschelle, 1993). It provides learners with an explanatory structure to communicate and sort out the world (Smith, 1991), and can act as a filter for new learning (Smith et al., 1993). Prior knowledge can also interfere with concept mastery, containing incorrect understandings or what have been most commonly labeled as "misconceptions" (Peşsman & Eryilmaz, 2010). Further, it has been found that these misconceptions are not altered by traditional methods of instruction (Laws, Sokoloff, & Thornton, 1999); they can persist (Smith et al., 1993).

Undergraduates in astronomy bring prior knowledge from personal experience of astronomical content to their coursework. However, students' encounters can be limited. For example, an understanding of astronomical distance originates with "… personally experienced distances in the everyday world, whereas most distances in astronomy are vastly larger" (Miller & Brewer, 2010, p. 1551). Confusion about astronomical

* **Acknowledgments:** This work was generously supported by the National Science Foundation through DUE-1140440.
Katharyn Ellen Ketter Nottis, Ph.D., professor, Department of Education, Bucknell University.
Edwin (Ned) Ladd, Ph.D., professor, Department of Physics and Astronomy, Bucknell University.
Alyssa Goodman, Ph.D., professor, Department of Astronomy, Harvard University.
Patricia Udomprasert, Ph.D., project director for the WorldWide Telescope Ambassadors Program, Harvard College Observatory, Harvard University.

distances has been found in previous research (e.g., Miller & Brewer, 2010; Trumper, 2001). Miller and Brewer (2010) discovered that the percentage of students underestimating astronomical distances increased as the distance of an object increased. For example, 33% of the students in their sample underestimated the distance to the Moon while 99% underestimated the distance to the closest galaxy. The problem did not seem to be one of scale as much as it was a change in scale. The researchers also noted that for about a third of the participants, the issue was not so much their estimation of astronomical distances as it was, "… underestimating the enormous leap in distance between objects in the solar system and between interstellar objects" (Miller & Brewer, 2010, p. 1557).

Another area where conceptual confusion has been found in astronomy is with the concept of scale, a concept that transcends disciplinary boundaries. The American Association for the Advancement of Science (AAAS) has noted that an understanding of scale is crucial to comprehending, "[T]he immense size of the cosmos, the minute size of molecules, and the enormous age of the earth" (AAAS, 1993, p. 276). In the absence of understanding scale, it is challenging to fully comprehend macroscopic and microscopic concepts as well as the large passage of time (Miller & Brewer, 2010).

In addition to distance and scale, misconceptions have also been seen in undergraduates' understanding of size and structure (Nottis & Ladd, 2014; Nottis, Ladd, Goodman, & Udomprasert, 2014). Table 1 lists some commonly found misconceptions.

Table 1

Common Misconceptions in Astronomical Size, Scale, and Structure

Content area	Misconception
Size	1. Students think that planets and stars have similar sizes. 2. Students think that all stars have the same size and luminosity. 3. Students underestimate the physical size of astrophysical objects. 4. Students confuse the size of the Milky Way Galaxy with the size of the universe.
Scale	1. Students think that the distance to the nearest stars is only a few times larger than the size of our Solar System. 2. Students underestimate the distances between galaxies.
Structure	1. Students think that the Earth, Solar System, and/or nearby stars are not part of the Milky Way Galaxy. 2. Students think that most objects outside our Solar System are approximately the same distance from Earth. 3. Students underestimate the aspect ratio of the Milky Way Galaxy's stellar distribution. 4. Students think that the distances between galaxies are not changing with time.

Note. Source: Personal conversation (Ladd, 2013).

Petcovic and Ruhf (2008) have noted that teachers need to know students' preconceptions in order to determine whether their conceptual understandings have altered by instruction. Such assessments are also helpful for instructors wanting to help students' construct conceptual understanding (Anderson, Fisher, & Norman, 2002) and to determine if misconceptions have altered (Shallcross, 2010). However, it is important that these evaluations focus on meaningful learning of science content or assess conceptual understanding rather than memorization of facts and formulas (Bransford, Brown, & Cocking, 2000; Lightman & Sadler, 1993).

There is a need for an assessment that can evaluate students' conceptual understandings of size, scale, and structure, and detect misconceptions. In order to ensure that concepts are assessed, test developers need to be aware that asking questions where students could have memorized a specific distance (e.g., Earth to the Sun) may be a measure of memory rather than conceptual understanding of distance (Miller & Brewer, 2010). Further, such evaluations may not detect students' misconceptions in astronomy. Another issue in astronomy is

a tendency to order distracters on multiple-choice tests sequentially with the largest (and correct) choice as the last one (Miller & Brewer, 2010). This pattern is quickly detected by test-savvy students. In addition, such an arrangement prevents students from overestimating, thereby failing to give them the opportunity to demonstrate whether they would over-rate an astronomical distance. There is a need to include questions with distracters larger than the correct answer in astronomical assessments.

One promising way to evaluate students' understanding of size, scale, and structure is to use concept inventories.

Concept Inventories

Concept inventories are multiple-choice assessments modeled on the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992). They are designed to assess conceptual understanding rather than students' ability to recall factual information. Prior research (e.g., Wuttiprom, Sharma, Johnston, Chitaree, & Soankwan, 2009) has demonstrated that concept inventories have a number of advantages including easy administration, objective scoring, and the ability to be analyzed with statistics. In addition, students' understanding of astronomical distances has also been previously measured using multiple-choice tests (e.g., Sadler, 1992).

As pre-tests used to determine students' prior knowledge, concept inventories have revealed differences among students in the same physics classes (Kost, Pollack, & Finkelstein, 2007; Noack, Antimirova, & Milner-Bolotin, 2009). Even more disturbing, some research has found that discrepancies between those who scored higher and those who scored lower on pre-tests were maintained throughout an entire course (Noack et al., 2009). Concept inventories can also be used to quantify students' increased understanding and the degree to which any naïve conceptions have been altered by comparing pre- with post- test results (Shallcross, 2010). The use of pre-post testing with concept inventories has revealed that some misconceptions do not alter after instruction.

Purpose of the Study

The purpose of the current study was to begin the process of developing a reliable and valid Size, Scale, and Structure Concept Inventory (S3CI) that could assess undergraduates' understanding of size, scale, and structure prior to and after instruction in introductory astronomy courses, detect the presence of previously documented misconceptions, and evaluate instructional effectiveness. While concept inventories have been used to assess astronomical concepts (e.g., Sadler, Coyle, Miller, Cook-Smith, Dussault, & Gould, 2010), none was found that specifically targeted these concepts. Although single questions have been incorporated into existing inventories (e.g., Astronomy and Space Science Test) (Sadler et al., 2010), adequate evaluation of students' understandings of size, scale, and structure as well as the detection of misconceptions needs a concept inventory specifically focused on those concepts.

For the purposes of question development, size was operationally defined as an understanding of the relative and absolute dimensions of astrophysical objects, such as stars, galaxies, and the universe itself. Scale was defined as an understanding of the distances between objects in the universe, including the tremendously large differences between distances within a structure (e.g., the Solar System) and between structures (e.g., to the nearest star). Finally, structure was defined as an understanding of the geometry of astrophysical objects, as well as their relative positions in the universe, both hierarchically (e.g., stars are located within galaxies), and in distance from Earth. Additionally, it included an understanding of how these objects' positions evolve in time.

There were also two secondary purposes of this study: to evaluate instructional effectiveness by comparing pre- with post-test results and to determine the impact of a newly developed lab on parallax, the first in a suite of labs enhanced with activities using the WorldWide Telescope (WWT). For this pilot study, students' understandings of parallax were assessed with a sub-set of six questions specifically targeting parallax taken from the beta version of the S3CI.

Methodology

Design and Procedure

The S3CI is being developed through what previous researchers have labeled as an "iterative process" (e.g., Petcovic & Ruhf, 2008, p. 253; Wuttiprom et al., 2009, p. 644). The following iterative steps suggested by Richardson (2004) were used to start its development:

1. Determine concepts to be included.
2. Study student learning process regarding those concepts.
3. Construct several multiple-choice questions for each concept.
4. Administer beta version of the inventory (24 questions in the current study) to as many students as possible. Examine reliability. In this case, estimates of internal consistency reliability were determined using the Kuder-Richardson Formula #20 (KR20).
5. Revise the inventory to improve readability, validity, and reliability.

Content experts in astronomy developed the initial version of the instrument using misconceptions about astronomical size, scale, and structure identified from past student work in introductory astronomy classes. The researchers developed and incorporated questions addressing each misconception in this initial concept inventory.

The researchers used both numerical and analogical questions to determine students' understanding and attempted to arrange distracters so that the numerically largest option was not always last. When possible, a numerical question was paired with a non-numerical question to assess understanding of the same concept. An example of paired questions from the beta version can be found in Appendix A.

Other experts in astronomy reviewed the instrument and provided feedback on whether they felt a question addressed a designated concept area and if the distracters were appropriate. Ways to improve the questions were also solicited. Undergraduate students in an introductory astronomy course were then given the beta version of 24 multiple-choice questions as both a pre- and post- test. Pre-tests were given in the first couple weeks of class while post-tests were administered in the last two weeks. Estimates of internal consistency reliability, a measure of the consistency of individual or subsets of questions across an instrument, were then determined (Huck & Cormier, 1996). The KR20 was used because it was assumed that items had varied levels of difficulty (Fraenkel, Wallen, & Hyun, 2012). Researchers aimed for estimates of internal consistency reliability at or greater than 0.70, a standard when assessments are used for research purposes (Fraenkel et al., 2012).

Researchers developing multiple-choice assessments in higher education have tended to use Classical Item Analysis (Libarkin & Anderson, 2006). Classical theory comes from the work of Gullikensen (1950) and has been more recently labeled as Classical Reliability theory or True Score theory (Suen, 1990). In this theory, information about identifiable factors of individual test questions provide a guide for "… the improvement of the test, and thus maximize the ultimate reliability of the total score" (Suen, 1990, p. 71).

As part of the initial development of the S3CI, a conventional item analysis was conducted, guided by classical theory. Two characteristics of the individual test items were examined to improve the reliability of the total score: item difficulty or the percentage getting an item correct and item discrimination or the difference between the proportions of high- and low- achieving students getting an item correct. Item discrimination was calculated by comparing the top third of the students with the bottom third (Patten, 2001). Poor discriminators were considered questions with a level below 0.20, as recommended by Brown (1983). The results of the item analysis guided recommendations for the revisions of this first version of the assessment.

A secondary purpose of this study was to compare pre- and post- test results on the entire concept inventory and a sub-set of questions addressing parallax measurements. A pre-experimental, one group pre-test-post-test design (Huck & Cormier, 1996) was used for this part of the study. Descriptive statistics examined changes in knowledge, as measured by the overall scores of participants on the concept inventory. Paired samples *t*-tests were used for significance testing. Effect sizes were also calculated to evaluate "the magnitude of a difference between the means of two groups …" (Fraenkel et al., 2012, p. 248). Since *t*-tests were used, Cohen's *d* was calculated as a measure of effect size.

Participants

In Fall 2013, the first version of the S3CI was piloted with a sample of convenience in an introductory astronomy course at a small private liberal arts institution in the northeastern United States. There were 37 who completed the pre-test and 41 who took the post-test. The participants included both majors and non-majors. Almost two thirds were male, 77.8% listed their race as White, and 44.4% were first-year students. The remainder was: 31.1% sophomores, 15.6% juniors, and 8.9% seniors. While 75.6% had a high school physics course, only 8.9% took Advanced Placement (AP) Physics. Just 4.4% had taken a course in astronomy in high school.

Parallax Measurements Lab

The general concept of astronomical parallax is usually described as the apparent change in the direction to an object when viewed from two different vantage points. The term "parallax angle" (Bennett, Donahue, Schneider, & Voit, 2014) has a specific connotation in astronomy. It means the change in apparent direction to a celestial object when viewed from two vantage points separated by one astronomical unit (AU); one AU is the average distance between the Earth and Sun. Parallax cuts across concepts but primarily addresses the matter of "size" in the universe.

Also, as part of this study, a new lab on Parallax Measurements using the WWT was developed (Ladd, Gingrich, Nottis, Udomprasert, & Goodman, 2014), based on an existing non-WWT parallax lab written by faculty members of the Department of Physics and Astronomy at the institution where the study was conducted. The WWT[1] is described on the Website as:

> … A rich visualization environment that functions as a virtual telescope, bringing together imagery from the best ground- and space-based telescopes to enable seamless, guided explorations of the universe. WWT, created with Microsoft's high-performance Visual Experience Engine, enables seamless panning and zooming across the night sky blending terabytes of images, data and stories from multiple sources over the Internet into a media-rich, immersive experience. The WWT experience scales from a Web browser all the way to multi-channel full dome in some of the world's most advanced planetariums. (http://worldwidetelescope.org/About/FAQ)

[1] The WWT can be downloaded for free at http://www.worldwidetelescope.org.

The lab begins with a WWT "tour," an activity that guides students through the concepts of astronomical parallax with frequent pauses for interactive exploration. Tour components involve an exploration of the three-dimensional distribution of the stars in Orion and the Big Dipper, an examination of how the Big Dipper stars appear to change positions when viewed from a non-Earthbound perspective, and an opportunity for students to directly measure the changes in the angular separation between Big Dipper stars as the observing perspective changes. After the "tour," students then work on understanding parallax in an outside area on campus, using lamp posts. The measurement methods are the same as those used for measuring distances to the stars but instead of astronomical telescopes, surveyor transits are used.

To determine the impact of the new Parallax Measurements Lab, six questions from the S3CI targeting relevant concepts were treated as a sub-test and used as both a parallax pre- and post- test.

Results

Reliability and Item Analysis

As can be seen in Table 2, the estimate of internal consistency reliability was lower than desired in this first version of the S3CI.

Table 2

Estimates of Internal Consistency Reliability on the Beta Version of the S3CI

Measure of reliability	Pre-test	Post-test
Split-half (odd-even) correlation	0.48	0.46
KR20	0.54	0.46

An examination of the difficulty and discrimination of each question, as shown in Table 3, revealed some problematic questions in need of revision and/or elimination. Half of the questions were considered poor discriminators according to Brown's (1983) recommendation. Five of these failed to discriminate at all (0.00). In addition, some questions had distracters selected by only a few or no participants.

Table 3

Beta Post-Test Item Difficulty and Discrimination

Question	Difficulty	Discrimination index
1	7.3	0.07
2	43.9	0.14
3	31.7	0.00
4	70.0	0.14
5	39.0	0.64
6	82.9	0.43
7	41.5	0.43
8	46.3	0.71
9	87.8	0.21
10	75.6	0.14
11	90.2	0.14
12	56.1	0.64
13	26.8	0.00
14	31.7	0.00
15	65.9	0.00
16	73.2	0.29

(Table 3 to be continued)

17	12.5	-0.07
18	42.5	0.29
19	75.6	0.43
20	100.0	0.00
21	85.4	0.43
22	19.5	0.29
23	63.4	0.71
24	75.6	0.07

The most difficult question was Question #1 (see Appendix B), which also did not discriminate between high and low performers. More students got the question correct on the pre-test, which suggests a pattern of guessing. Figure 1 shows a graph of the pre- and post- test responses to this question. In these graphs, it is also clear that some distracters worked better than others. For example, only one student selected "a" on both the pre- and post- test.

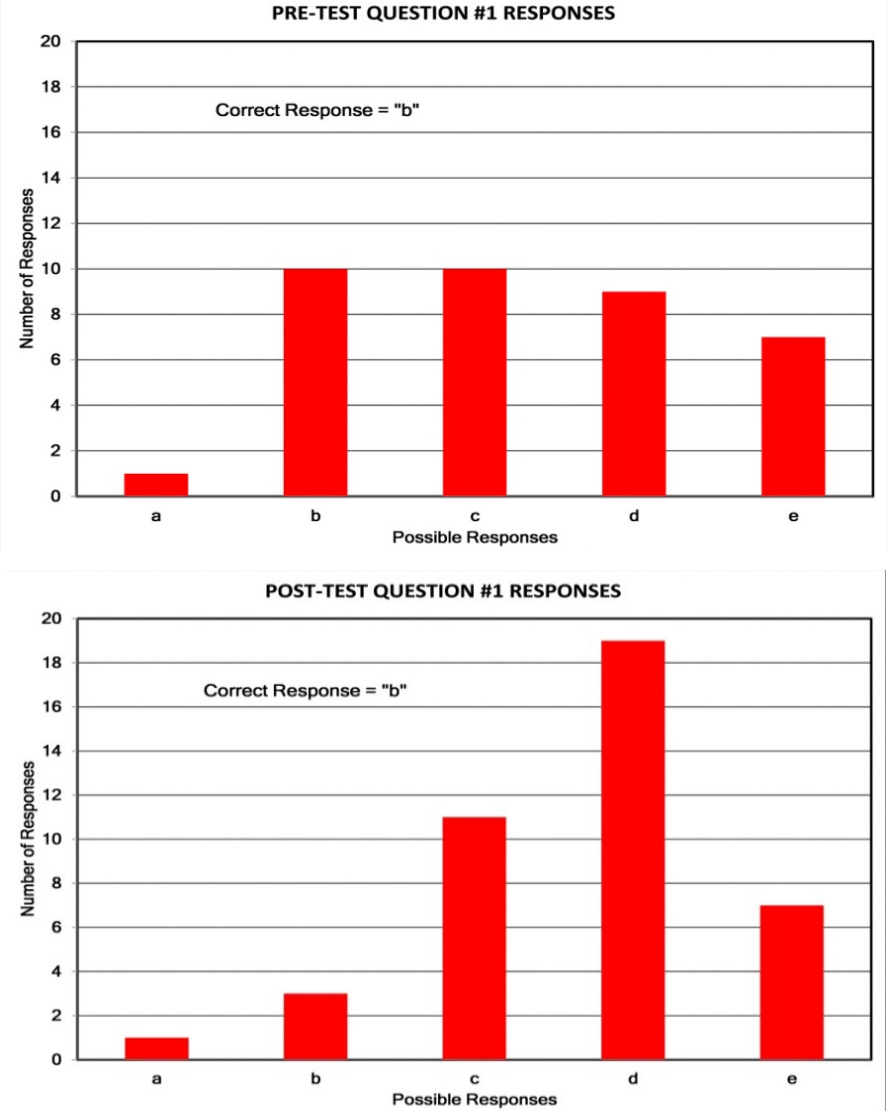

Figure 1. Comparison of students' pre- and post- test responses to Question #1.

Students' Understanding of Size, Scale, and Structure

The mean pre-test score was 11.24 points (46.8%) and the mean post-test score was 13.41 points (55.9%). Paired samples *t*-tests revealed that there was a significant improvement from pre- to post- test on the S3CI with a moderate effect size ($t(33) = -2.80$; $p < 0.01$; $d = 0.48$).

Students' Understanding of Parallax

To determine the impact of the new lab, six questions targeting relevant concepts were treated as a sub-test and also examined separately. The estimates of internal consistency reliability for the post-test were higher than the beta version of the entire instrument: Split-half was 0.41 and KR20 was 0.61. Table 4 shows that four of the six questions showed improvement from pre- to post- test. Question #22 (see Appendix C) remained very difficult. Paired samples *t*-test showed that there were no significant differences from pre to post on the parallax sub-test.

Table 4

Impact of the Parallax Lab on Targeted Questions

Question #	Percentage of participants getting question correct on pre-test (%)	Percentage of participants getting question correct on post-test (%)
5	38.9	39.0
6	64.9	82.9
7	29.7	41.5
8	36.1	46.3
22	18.9	19.5
23	51.4	63.4

Conclusions and Educational Implications

Misconceptions resistant to change through traditional teaching methods are of particular interest to science educators, especially those that can impact understanding in multiple disciplines. However, there is a need for reliable and valid instruments to assess conceptual understanding and whether new pedagogies and technologies have altered those misconceptions.

This pilot study initiated the development of the S3CI and revealed that there were a number of issues in this initial version. First, estimates of internal consistency reliability were lower than the recommended standard of 0.70 for research purposes (Fraenkel et al., 2012). An item analysis showed that some questions did not discriminate well between high and low performers. In addition, some distracters did not seem plausible to respondents, even on the pre-test. This could be seen with Question #1, the most difficult question on the post-test, where only one person selected one of the distracters on both the pre- and post- test. Future versions of the S3CI should remove distracters not viewed as credible by students.

Some questions also had response patterns suggestive of instruction inadvertently prompting students to overgeneralize the idea that astronomical distances are greater than believed. For example, the pre-test responses to Question #1 appeared to show a pattern of random distracter selection, with the exception of distracter "a," which was an unappealing choice to participants. However, after a semester of instruction where students learned how great astronomical distances are, they tended to select distracters with a very large size ("c" or "d"), but not necessarily the largest or the correct response. Participants may have assumed the largest size was incorrect due to previously learned test-taking strategies. Further investigation is needed. However, before

conclusions can be drawn about instruction, questions need to be revised based upon the results of the item analysis and the reliability of the total instrument needs to be improved. There is an additional challenge of phrasing questions in ways that are plausible to non-majors and yet do not have discernible patterns that test-wise students can detect. All revisions should consider this issue as well.

An examination of targeted questions examining parallax measurements showed that four of the six questions had good gains although there were no significant differences in students' understanding from pre to post on the entire group of questions. Two questions showed no improvement in understanding. One of these (#22) was very difficult and did not discriminate well. It asked how the night sky would look if viewed from Pluto. Given the way this question is posed, students needed prior knowledge about the relative distances between the Earth, Pluto, and nearby stars so more than parallax was being assessed. Students did not seem to understand that relative to the stars and constellations, the distance between the Earth and Pluto is miniscule. Respondents appeared to think that Pluto is much closer to the stars than the Earth, so as a result, this question may have tested knowledge of scale as much as an understanding of parallax. Students also seemed to think that Pluto, when compared with Earth, must be different in some way. Their confusion with this may reflect a basic understanding of parallax (i.e., that the view should look different from different perspectives) without a more nuanced understanding of the magnitude (and therefore observability) of the effect. This was possibly due to what was learned about parallax in the new lab. It is also conceivable, as noted previously, that the confusion was tied more to their understanding of scale and distance rather than parallax. An additional issue with this question was that the correct response has different wording from the other distracters. This should be altered.

The method of instruction used in the parallax lab may have also contributed to some of the issues with students' responses. An existing parallax lab was used and then adapted. The lab began with a tutorial of the WWT where students take a "tour" of the universe and learn about parallax. That part of the lab was followed by a hands-on look at parallax on campus. While it could be argued that this enhanced lab offers a balance of computer-based and on-campus activities, the connection between the two segments may not be as clear to students; there could be a separation in their "real world" and "universe" models of parallax, which impacted students' understanding. In addition, no questions on the beta version of the S3CI addressed parallax as it was taught on campus.

Although computer visualization and simulation tools may help students increase their understanding of these difficult concepts, where and how the WWT is used in the lab needs to be carefully considered. Previous science education research has investigated the placement of computer visualization and simulation tools in the learning process (e.g., Brant, Hooper, & Sugrue, 1991; Carlson & Andre, 1992; Hargrave & Kenton, 2000; Udomprasert et al., 2015) and how visualization may need to be altered for certain groups of students (e.g., Lin, 2001). For example, Hargrave and Kenton (2000) noted the benefits of using simulations as pre-instruction including showing ongoing natural processes and phenomena thereby setting the stage for what would be considered more traditional instruction, and giving students' autonomy over the course content and their learning. Lin (2001) examined the effect of presentation format (text, still graphics, and animation) and students' prior knowledge (novice and experienced) on their descriptive and procedural knowledge in physics. Descriptive knowledge was operationally defined as more factual learning while procedural knowledge was described as an understanding of problem-solving steps or procedures used in physics. The researcher found a significant difference in descriptive learning with prior knowledge and a significant interaction between prior knowledge and format. With procedural knowledge, Lin (2001) did an analysis of covariance (ANCOVA)

controlling for previous physics and mathematics scores. There were significant effects for format and prior knowledge as well as a significant interaction. The addition of animation did not improve the learning of novices in that study. The researcher found that "… The need for visual format differs when there is knowledge discrepancy among learners" (Lin, 2001, p. 146). Given that only 4.4% of the current sample had taken a course in astronomy in high school, there may be a need to consider when the WWT additions are used and the way the computer visualization is presented, so optimal use can be facilitated.

More recently, Udomprasert, Goodman, Zhang, et al. (2014) and Udomprasert, Goodman, Sadler, et al. (2015) designed a middle school lab experience to help students understand the cause of the Moon's phases, using a combination of physical models (Styrofoam balls and lamps) and computer models (WWT). The researchers tested how model order (Foam then WWT *vs.* WWT then Foam) would impact student learning. Students in their study who used WWT first had stronger learning gains and explained Moon phases using more sophisticated ideas on the post-assessment than those who used the Styrofoam ball/lamp model first. One reason could be that interacting with WWT provided students with a better foundation for understanding how to use and manipulate the foam ball model effectively.

There were a number of limitations in this pilot study. First, a sample of convenience from one institution of higher education and from one astronomy course took the first version of the S3CI. Future versions of the S3CI should be given to a larger number of participants from more than one institution. Second, Classical Reliability theory has major limitations including the fact that both item difficulty and discrimination are dependent upon the participants (Fan, 1998). While care was taken to obtain a sample that would be similar to the group of students who might use the inventory in the future, the sample was small and from one university.

The development of the S3CI is just beginning and shows promise in documenting conceptual understanding of difficult astronomical concepts as well as the impact of new technology in the learning process. Future work should include an evaluation of distracters, removing and revising questions with low or negative discrimination indices in an effort to raise the internal consistency reliability, and obtaining qualitative data from open-ended questions where students explain why they selected a specific response. These responses can guide further distracter development and ultimately lead to a reliable and valid assessment. Simultaneously, the development of astronomy labs using the WWT should consider when and how the inclusion of this virtual telescope can most benefit students' understanding.

References

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy: Project 2061*. New York, N.Y.: Oxford University Press.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching, 39*(10), 952-978.

Bennett, J., Donahue, M., Schneider, N., & Voit, M. (2014). *Cosmic perspective: Stars, galaxies, and cosmology* (7th ed., p. 489). Pearson Higher Education Publishing.

Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience and school.* Washington, D.C.: National Academy Press.

Brant, G., Hooper, E., & Sugrue, B. (1991). Which comes first the simulation or the lecture? *Journal of Educational Computing Research, 7*, 469-481.

Brown, F. G. (1983). *Principles of educational and psychological testing* (2nd ed.). New York, N.Y.: Holt, Reinhart, & Winston.

Carlson, D. D., & Andre, T. (1992). Use of microcomputer simulation and conceptual change text to overcome student preconceptions about electrical circuits. *Journal of Computer Based Instruction, 19*(4), 105-109.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357(25).

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York, N.Y.: McGraw-Hill.

Gullikensen, H. (1950). *Theory of mental tests*. New York, N.Y.: Wiley.

Hargrave, C. P., & Kenton, J. M. (2000). Pre-instructional simulations: Implications for science classroom teaching. *The Journal of Computers in Mathematics and Science Teaching, 19*(1), 47-58.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141-158.

Huck, S. W., & Cormier, W. H. (1996). *Reading statistics and research* (2nd ed.). New York, N.Y.: Harper Collins College Publishers.

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2007). Investigating the source of the gender gap in introductory physics. *Physics Education Research Conference, 951*, 136-139.

Ladd, E. F., Gingrich, E. C., Nottis, K. E. K., Udomprasert, P., & Goodman, A. A. (2014, August). Combining real world experience and WorldWide Telescope visualization to build a better parallax lab. Poster presented at *The Astronomical Society of the Pacific Conference*, San Francisco, C.A..

Laws, P., Sokoloff, D., & Thornton, R. (1999). Promoting active learning using the results of physics education research. *UniServe Science News,* 13.

Libarkin, J. C., & Anderson, S. W. (2006). The Geoscience Concept Inventory: Application of Rasch analysis to concept inventory development in higher education. In *Applications of Rasch measurement in science education* (pp. 45-73). Maple Grove, M.N.: JAM Press.

Lightman, A., & Sadler, P. (1993). Teacher predictions versus actual student gains. *The Physics Teacher, 31*, 162.

Lin, C. (2001). Formats and prior knowledge on learning in a computer-based lesson. *Journal of Computer Assisted Learning, 17*, 409-419.

Miller, B. W., & Brewer, W. F. (2010). Misconceptions of astronomical distances. *International Journal of Science Education, 32*(12), 1549-1560.

Noack, A., Antimirova, T., & Milner-Bolotin, M. (2009). Student diversity and the persistence of gender effects on conceptual physics learning. *Canadian Journal of Physics, 87*, 1269-1274. doi:10.1139/P09-108

Nottis, K. E. K., Ladd, N., Goodman, A., & Udomprasert, P. (2014, October). Preliminary development of an instrument to assess size, scale, and structure concepts in introductory astronomy. Paper presented at *The Northeastern Educational Research Association Conference*, Trumbull, C.T..

Nottis, K., & Ladd, N. (2014, January). Development of an instrument to assess size, scale, and structure concepts in introductory astronomy. Paper presented at *The Hawaii International Conference on Education*, Honolulu, Hawaii.

Patten, M. L. (2001). *Questionnaire research: A practical guide* (2nd ed.). Los Angeles, C.A.: Pyrczak.

Peşsman, H., & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research, 103*(3), 208-222.

Petcovic, H. L., & Ruhf, R. J. (2008). Geoscience conceptual knowledge of preservice elementary teachers: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education, 56*(3), 251-260.

Richardson, J. (2004). Concept inventories: Tools for uncovering STEM students' misconceptions. In *Invention and impact: Building excellence in undergraduate science, technology, engineering and mathematics (STEM) education.* Washington, D.C.: AAAS. Retrieved from http://www.aaas.org/publications/books_reports/CCLI/

Sadler, P. M. (1992). The initial knowledge state of high school astronomy students. *Dissertation Abstracts International, A53/05*, 1470.

Sadler, P. M., Coyle, H., Miller, J., Cook-Smith, N., Dussault, M., & Gould, R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 national science standards. *Astronomy Education Review, 8*(1).

Shallcross, D. C. (2010). A concept inventory for material and energy balances. *Education for Chemical Engineers, 5*, e1-e12.

Shuell, T. J. (1992). Designing instructional computing systems for meaningful learning. In M. Jones, & P. H. Winne (Eds.), *Adaptive learning environments: Foundations and frontiers* (pp. 19-54). New York, N.Y.: Springer-Verlag.

Smith, E. L. (1991). A conceptual change model of learning science. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), *The psychology of learning science* (pp. 43-63). Hillsdale, N.J.: Lawrence Erlbaum.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Suen, H. K. (1990). *Principles of test theories*. Hillside, N.J.: Lawrence Erlbaum.

Trumper, R. (2001). Assessing students' basic astronomy conceptions from junior high school through university. *Australian Science Teachers Journal, 41*, 21-31.

Udomprasert, P., Goodman, A., Sadler, P., Johnson, E., Lotridge, E., Jackson, J., … Trouille, L. (2015, April). Optimal model-order for a moon phases lab with virtual and physical components. Paper presented at *The American Educational Research Association*, Chicago, I.L..

Udomprasert, P., Goodman, A., Zhang, Z. H., Sunbury, S., Sadler, P., Dussault, M., … Constantin, A. M. (2014). Visualizing three-dimensional spatial relationships in virtual and physical astronomy environments. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, … L. D'Amico (Eds.), *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014* (Vol. 3). Boulder, C.O.: International Society of the Learning Sciences.

Wuttiprom, S., Sharma, M. D., Johnston, I. D., Chiaree, R., & Soankwan, C. (2009). Development and use of a conceptual survey in introductory quantum physics. *International Journal of Science Education, 31*(5), 631-654.

Appendix A: Numerical and Non-numerical Question Pair

The nearest stars (not including the Sun) are about ____ farther away than the Sun from Earth.

(a) 40 times

(b) 4,000 times

(c) 400,000 times

(d) 4,000,000 times

Using a scale model where the Earth is a ball about the size of your hand (5 inches or 12 cm.), about how far away would you have to put the nearest star outside our solar system?

(a) across a football field/soccer pitch

(b) across town

(c) a one-hour airplane flight

(d) on the other side of the Earth

(e) on the Moon

Appendix B: Question #1

The Sun is about _____ farther away from the Earth than the Moon is from the Earth.

(a) 40 times

(b) 400 times

(c) 4,000 times

(d) 40,000 times

(e) 400,000 times

Appendix C: Question #22

How would the night sky look to you if you observed it from the surface of Pluto?

(a) The stars and constellations would look the same as they do from Earth.

(b) The constellations would look different, but the brightness of the stars would be the same.

(c) The constellations would look different, and the stars would be brighter.

(d) The constellations would look different, and the stars would be fainter.