A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

2013, AER, 12(1), 010107, http://dx.doi.org/10.3847/AER2012045

Development of the Newtonian Gravity Concept Inventory

Kathryn E. Williamson
Montana State University, Bozeman, Montana 59717
Shannon Willoughby
Montana State University, Bozeman, Montana 59717
Edward E. Prather
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
Received: 12/5/12, Accepted: 05/7/13, Published: 07/2/13

Abstract

We introduce the Newtonian Gravity Concept Inventory (NGCI), a 26-item multiple-choice instrument to assess introductory general education college astronomy ("Astro 101") student understanding of Newtonian gravity. This paper describes the development of the NGCI through four phases: Planning, Construction, Quantitative Analysis, and Validation. We discuss the evolution of the instrument through three versions, including the refinement of a set of four concept domains and nine examples of items to illustrate how expert review, student interviews, and Classical Test Theory statistics informed our approach. We conclude that the NGCI is a reliable and valid instrument.

1. INTRODUCTION

In this paper, we provide a detailed discussion about the development of a new concept inventory that is designed to assess student understanding of Newtonian gravity in general education introductory college astronomy courses, hereafter "Astro 101." The Newtonian Gravity Concept Inventory (NGCI) probes several of the common student conceptual and reasoning difficulties with the concept of gravity as described in the literature (Ameh 1987; Smith and Treagust 1988; Sharma *et al*. 2004; Dostal 2005; Kavanagh and Sneider 2006–2007a and 2006–2007b; Feeley 2007; Asghar and Libarkin 2010; Williamson and Willoughby 2012). Specifically, the concept mapping, phenomenographic analysis of student-supplied responses to open-response questionnaires, and characterization of students' mental models described in our previous paper, Williamson and Willoughby (2012), guided the conceptual focus and item construction of the NGCI. This foundational work represents the Planning phase of the instrument development model of Benson and Clark (1982).

This paper discusses our approach to the three subsequent phases of Benson and Clark's (1982) model—Construction, Quantitative Evaluation, and Validation—to provide a full disclosure of the research methods and results used to develop and refine the NGCI iteratively through three versions. Section 2 describes the use of Concept Inventories in Astro 101 courses and recaps the conclusions drawn from Williamson and Willoughby (2012) to ground the development of the NGCI. The Construction phase follows in Section 3, including the refinement of the four concept domains of the NGCI to bound the ideas (both scientifically correct and incorrect) assessed by the instrument. Here, we also outline our approach to item construction. The Quantitative Evaluation of pilot testing data of the three versions of the NGCI and the Classical Test Theory statistical analysis is provided in Section 4 to motivate changes throughout the multi-step development of the NGCI. In Section 5, we highlight the evolution of nine items to illustrate the iterative process by which multiple-choice questions were evaluated and modified to ensure the conceptual breadth, scientific accuracy, and item clarity of the NGCI. We draw on student and expert performance on the NGCI, as well as student interviews and expert review in Section 6 to argue for the validity of the instrument in measuring Astro 101 students' understanding of Newtonian gravity. Section 7 concludes with a summary and future plans.

2. PHASE ONE: PLANNING

In this section, we provide background to motivate our discussion of the development of the NGCI. First, we discuss the precedent and proven usefulness of Concept Inventories in the Astronomy Education Research community, and second, we review the foundational work in Williamson and Willoughby (2012) on characterizing student understanding of Newtonian gravity through student alternative models and misapplications of the scientific model of Newtonian gravity.

2.1 Concept Inventories

Multiple-choice Concept Inventories (CIs) have become important tools in the Astronomy Education Research community for assessing student learning and the effects of instructional interventions. CIs focus on a narrow domain of topics that are central or foundational to the overall Astro 101 curriculum (Bailey 2009). The multiple-choice questions used in a Concept Inventory endeavor to model students' natural language, minimize scientific jargon, and provide research-based distractor choices that represent common student naïve beliefs and reasoning difficulties. There are currently other CIs for Astro 101 courses that measure student understanding about lunar phases (Lindell and Olsen 2002), star properties (Bailey 2006), the green house effect (Keller 2006), light and spectroscopy (Bardar *et al*. 2007), and the solar system (Hornstein *et al*. 2011). In addition to these topics, the topic of Newtonian gravity is a foundational topic for Astro 101. An understanding of gravity enhances students' ability to solve problems and reason about the motion of bodies in space, the formation of planets and stars, and the large-scale structure of the universe. To quickly and reliably probe Astro 101 students' understanding of this important topic, and to provide a useful tool for faculty to assess success in their classes, we have developed the Newtonian Gravity Concept Inventory (NGCI).

2.2 Common Student Difficulties With Newtonian Gravity

Building on studies that explored children's understanding of gravity (see Kavanagh and Sneider 2007a and 2007b for a review) and physics students' understanding of gravity (for example, Sharma *et al*. 2004 and Dostal 2005), Williamson and Willoughby (2012) implemented a grounded theory method (Creswell 2007) and phenomenographic analysis of Astro 101 student-supplied responses to twenty-three open-ended questions about gravity. This process was essential to uncovering the breadth of Astro 101 student ideas related to gravity. In this section, we recap the main student alternative models and misapplications of the scientific model of Newtonian gravity.

Williamson and Willoughby (2012) describe three main alternative models Astro 101 students implement when thinking about gravity—The Boundary Model, the Orbital Indicator Model, and the Mixing of Forces Model. The Boundary Model encapsulates student misconceptions related to a sudden change in the gravitational force for objects leaving a celestial body. For example, the majority of students believe that one simply must leave the Earth (and its atmosphere) to feel zero gravitational force from Earth. These students likely view the apparent weightlessness of astronauts as zero gravitational force and associate weightlessness with the absence of an atmosphere. Therefore, the atmosphere could be viewed as either a boundary that contains Earth's gravity or simply an indicator of the presence of gravity. The Orbital Indicator Model is a related idea in which the objects that orbit a planet indicate the strength of gravity on the planet's surface. Students apply this in two opposing ways: (1) A planet must exert a strong gravitational force on objects if it can hold far away massive objects in orbit, or (2) A planet must exert a strong gravitational force on objects if they are orbiting close by because it essentially "reigns" them in. The third alternative model, the Mixing of Forces Model, explains the well-documented (Smith and Treagust 1988; Sharma 2004; Feeley 2007; Asghar and Libarkin 2010) student misconception that gravity is confounded with other forces associated with magnetism, rotation, and atmospheric pressure.

Williamson and Willoughby (2012) found that while students may hold a combination of these alternative models, they may also simultaneously hold other misconceptions as well as correct conceptions related to the scientific model. Students interviewed in this study who had even a cursory knowledge of Newtonian gravity typically understood that mass and distance are important factors in determining gravitational force, but they have trouble applying these ideas in a consistent and coherent manner. For instance, some students believed that only very massive objects can exert a force on other objects, or they confused the ideas of mass and density. The most common errors associated with distance were in measurement, with students measuring the distance as the

radius of the planet or the distance from the surface. Additionally, it was clear that many students do not think of gravity as a mutual force of attraction, but rather as a property that emanates from a material or object.

One may refer to Williamson and Willoughby (2012) for a more in depth discussion of how these misconceptions manifested in student responses to specific questions about gravity. For this paper, we provide this information to ground the Construction phase of the development process discussed in Section 3.

3. PHASE TWO: CONSTRUCTION

Constructing the NGCI involved the construction of (1) a set of four concept domains to define the aspects of gravity that the instrument is designed to measure, and (2) multiple-choice items that probe these concept domains, which are written in the natural language of students, and include effective distractor choices.

3.1 Concept Domain

The Planning work in Williamson and Willoughby (2012) outlined a concept map of four aspects of gravity that guided the development of open-ended questions—Direction, Force Law, Gravity Unchanged by Other Factors, and Orbital Effects. However, as items are developed and piloted with different groups, one can more clearly define the construct being measured (Wilson 2005), and these four aspects of gravity were refined into four different "Concept Domains"—Directionality, Force Law, Independence of Other Forces, and Threshold. Table 1 provides the breadth of ideas contained within these conceptual domains and probed by items on the NGCI. These ideas represent both correct and incorrect notions, including the most common student difficulties described in previous work (Ameh 1987; Smith and Treagust 1988; Sharma 2004; Dostal 2005; Kavanagh and Sneider 2006–2007a and 2006–2007b; Feeley 2007; Asghar and Libarkin 2010; Williamson and Willoughby 2012) and uncovered throughout the development process. In Section 5, we use the Classical Test Theory results presented in Section 4, along with expert review and student interviews, to discuss exactly how some of the major conceptual changes to the NGCI led to these four concept domains.

With these concept domains, one can see that a student who understands Newtonian gravity as measured by the NGCI should correctly and coherently reason about

- the direction of the gravitational force.
- the magnitude of the gravitational force.
- the relationship between mass and gravitational force.
- the relationship between distance and gravitational force.
- how the gravitational force is independent from other forces.
- how the gravitational force behaves in limiting cases.

3.2 Item Construction

In constructing the multiple-choice items for the NGCI, we referred to the best practices outlined in Haladyna, Downing, and Rodriguez (2002). Their 31 guidelines group into five categories: Content Concerns, Formatting Concerns, Style Concerns, Writing the Stem, and Writing the Choices. Of these guidelines, several were revisited repeatedly throughout development of the NGCI, such as focusing each item to reflect a specific concept, using typical reasoning errors of students for distractor choices (i.e., the most-common misconceptions uncovered from the open-response questions in Williamson and Willoughby (2012)), using simple vocabulary, minimizing reading load, and keeping choices similar in length, consistent in content and grammatical structure, and in a logical order when applicable.

4. PHASE THREE: QUANTITATIVE EVALUATION

The NGCI was developed iteratively over three versions. In this Section, we discuss the test population for each version and provide descriptive statistics and Classical Test Theory Statistics to perform a quantitative evaluation of the instrument.

Table 1. The final Newtonian Gravity Concept Inventory (NGCI) concept domains. The questions of the NGCI probe student ideas within these four domains: Directionality, Force Law, Independence of Other Forces, and Threshold. Ideas represent the range of student ideas uncovered throughout the development process, with those that are correct marked with an asterisk

Directionality domain

Multiple objects

- *Direction of the total force is determined by superposition.
- Direction of the total force is toward the larger object only.
- Direction of the total force is toward the closer object only.

Relative motion

- *Direction of the force is not determined by the direction of motion.
- Direction of the force is determined by the direction of motion.

For objects on the surface of a large body

- *Direction of the force is toward the center of mass.
- Direction of the force is determined by the direction of apparent weight.
- Direction of the force is always perpendicular to the surface.

Force Law Domain

Determination of magnitude

- *Determined by the force equation with mass and distance.
- Can be blocked or diminished by another massive object.
- Can be estimated by the apparent weight.

The role of distance

- *Distance is measured from the center of mass of an object.
- Distance is measured from the surface of an object.
- Distance is measured by the radius of an object.
- *Distance squared and force of gravity are inversely related.
- Distance and force of gravity are inversely related.
- Distance and force of gravity are related in another way.
- Distance and force of gravity are not related.

The role of mass

- *Both masses matter.
- Only the larger mass matters.
- *Mass and force of gravity are directly related
- Mass and force of gravity are related in another way.
- Mass and force of gravity are not related.

Effects of density

- *Changing density does not change the gravitational force experienced by an object in space.
- Changing density does change the gravitational force experienced by an object in space.
- *Changing density does change the gravitational force experienced by an object on the surface.
- Changing density does not change the gravitational force experienced by an object on the surface.

Independence of other forces domain

Air pressure

- *Gravity is not affected by air pressure.
- Gravity is affected by air pressure

Magnetism

- *Gravity is not affected by magnetism.
- Gravity is affected by magnetism.

Rotation

- *Gravity is not affected by rotation.
- Gravity is affected by rotation.

Threshold domain

Distance threshold

- *There is no distance for which the force of gravity suddenly stops.
- There is a distance for which the force of gravity suddenly stops.
- There is a distance for which the force of gravity becomes constant and nonzero.

Mass threshold

- *There is no minimum mass for an object to experience a gravitational force.
- There is a minimum mass for an object to experience a gravitational force.

Atmospheric threshold

- *No sudden shift in the gravitational force experienced by objects occurs at the edge of Earth's atmosphere.
- A sudden shift in the gravitational force experienced by objects does occur at the edge of Earth's atmosphere.

Orbital threshold

- *Gravitational force does not always cause relative motion between objects.
- Gravitational force always causes relative motion between objects.

4.1 Pilot Testing

Version 3 of the NGCI is the final and most widely tested version. Version 3 was piloted both pre- and post-instruction at four large state universities and two community colleges (CC's). This provided a total of 925 Astro 101 students who participated in the NGCI Version 3 pretest, and 743 who participated in the post-test, with a total population average pre-instruction score of 43.71% ($SD = 19.01$) and post-instruction score of 55.49% ($SD = 20.93$) (See Note-1). Table 2 shows how preinstruction and postinstruction descriptive statistics varied by institution, including number of students, average normalized gain, and effect size with confidence intervals. The data for the University of California Davis were separated into the Solar System course and the Stars and Galaxies course, since Newtonian gravity is taught in both courses but applied in varying degrees and in different contexts. Table 3 shows the accompanying Version 3 demographic data. These percentages were calculated as prepost averages of the students who responded to the demographic questions at the end of the NGCI. Note that students who reported an age of 17 or younger were eliminated from the sample.

4.2 Classical Test Theory Analysis

Classical Test Theory (CTT) is one of the most straightforward and often used statistical methods for evaluating multiple-choice instruments (Ding and Beichner 2009). While CTT is highly sample dependent (Hambleton and Jones 1993), it can offer valuable information about the reliability and item functioning of an instrument, and it was used with each version of the NGCI to inform the development. Here, we will briefly review the basics of CTT, and we refer readers interested in the theory to Crocker and Algina (1986) and to examples of how it has been applied to instrument development in the discipline of Astronomy Education Research (Bailey 2006; Wallace and Bailey 2010; and Schlingman *et al*. 2012).

In CTT, the Cronbach's α statistic measures the internal consistency, or reliability, of an instrument. Values range between 0 and 1 and are highest when the variance of items is small compared to the variance of total test scores, with values over 0.70 generally accepted as indicating a reliable instrument. P-values measure the difficulty of each item. An item's P-value is the proportion of students that got the item correct. Here, however, we report item difficulty as D, the proportion of students that got the item *incorrect* so that higher values indicate more difficult items. With this adaptation, difficulty values lower than 0.20 indicate items that might be too easy and values higher than 0.80 indicate items that might be too hard (Schlingman *et al*. 2012). The point-biserial index, r_{pb}, measures an item's ability to discriminate between high ability students and low ability students. A value over 0.30 generally indicates that students' scores on that item are well-correlated with their total test scores.

Each version of the NGCI showed an improvement in Cronbach's α reliability, with the final Version 3 post-test value at $\alpha = 0.84$. For comparison, post-test Cronbach's α is 0.75 for the Lunar Phases Concept Inventory (Lindell and Olsen 2002), 0.72 for the Star Properties Concept Inventory (Bailey 2006), and 0.78 for the Light and Spectroscopy Concept Inventory (Schlingman *et al*. 2012). Figures 1 and 2 show CTT item difficulty and item discrimination values, respectively, for questions that remain on Version 3 of the NGCI, including accompanying Version 1 and Version 2 values where appropriate. One will note that most CTT item statistics for

Table 2. Pilot site data for the final version of the Newtonian Gravity Concept Inventory (NGCI), Version 3

Institution		N	Mean %	SD	$\langle g \rangle$	Cohen's *d* (95% C.I.)
Montana State University	Pre	278	44.12	17.51	0.09	0.29
	Post	245	49.42	19.29		(0.12–0.46)
University of Arizona	Pre	287	40.59	17.26	0.24	0.77
	Post	237	54.79	19.83		(0.59–0.95)
Westchester CC	Pre	72	37.39	16.57	0.55	0.91
	Post	66	54.78	21.39		(0.56–1.26)
Youngstown State University	Pre	107	36.88	16.36	0.27	0.93
	Post	59	53.72	21.05		(0.59–1.26)
Truckee meadows CC	Pre	17	42.08	12.65	0.49	2.12
	Post	8	70.67	15.24		(1.03–3.06)
UC Davis solar system	Pre	47	59.00	20.24	0.06	0.13
	Post	33	61.54	19.37		(−0.32–0.57)
UC Davis stars and galaxies	Pre	117	54.65	22.67	0.35	0.75
	Post	95	70.66	19.46		(0.47–1.03)

the NGCI are within the conventionally accepted bounds. Two items were too difficult for the Version 3 pretest, while only one of these remained too difficult for the post-test. And, four items had low discriminatory power for the pretest, while only two of these remained low for the post-test. For example, Items 10 and 25 are the two most difficult items on the survey, however, Item 10 is a relatively good discriminator of student ability, while Item 25 is not. Even students who perform very well on the survey overall often incorrectly answer Item 25 (see Section 5.3 for a discussion of why Item 25 remains on the final version of the NGCI). In general, Version 3 postinstruction item difficulty values are lower than preinstruction values, while post-test discrimination values are higher.

Figures 1 and 2 also show that early versions of many items were more difficult and less discriminating than the final versions, indicating that revisions throughout the development process led to clear item improvement. In the Evolution of Questions and Concept section, Section 5, we provide examples of the specific changes made to the NGCI questions and conceptual focus throughout the development process that led to these observed improvements in CTT statistics.

Table 3. Newtonian Gravity Concept Inventory (NGCI) Version 3 demographic data

		%
Major	Business	24.0
	Education	9.9
	Humanities, Social Sciences, Arts	29.1
	Science, Engineering, Architecture	20.0
	Other	17.0
Previous Courses	0	57.1
	1	30.4
	2 or more	11.5
Gender	Male	51.0
	Female	49.0
Age	18–30	96.9
	Older than 30	3.1

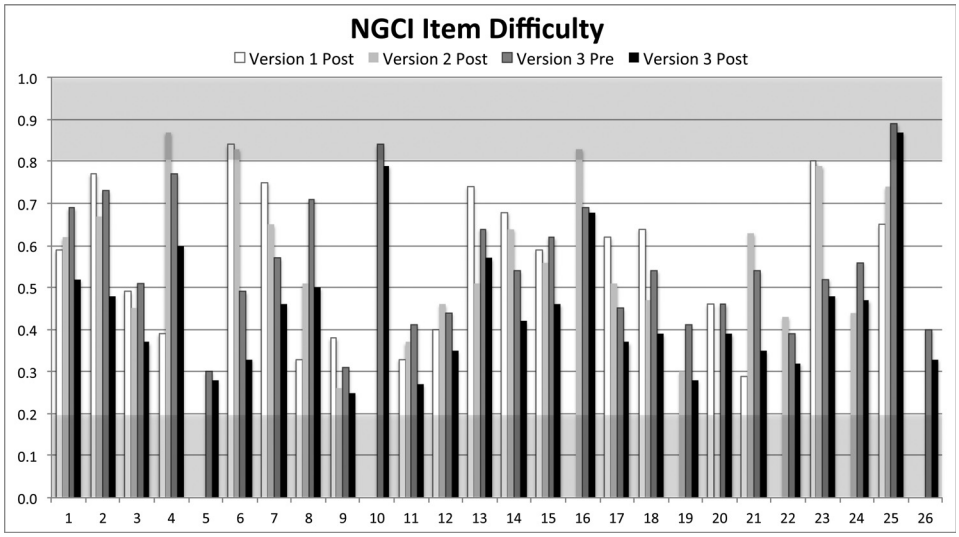

Figure 1. Version 3 NGCI item difficulty values, with accompanying values for Versions 1 and 2 where appropriate. Gray regions represent values that typically indicate an item is too difficult (greater than 0.80) or too easy (less than 0.20).

5. EVOLUTION OF QUESTIONS AND CONCEPTS

In this section, we explore the details of our choices in the genesis, inclusion, exclusion, and evolution of nine items. These are concrete examples of how we implemented the best practices for instrument development by including the results of the quantitative analysis presented in Section 4 as well as suggestions from expert review and student interviews. This discussion is organized according to the conceptual domains outlined in Section 3. Not only does this elucidate how these domains were refined throughout the development process, it grounds our assertion that the intended focus of the NGCI manifests in specific items.

5.1 Ontological Categories

Before we investigate the four concept domains of the NGCI, we first must discuss a significant change to the NGCI that occurred from Version 1 to Version 2. In particular, expert review of Version 1 indicated a need for ontological consistency in the language of the NGCI. This meant redesigning questions to refer to gravity as a force between two objects, rather than as *stuff* a physical entity has or as a field. While experts may reason about gravity using the ideas of both force and field. Gupta, Hammer and Redish (2010); Chi, Slotta, and de Leeuw

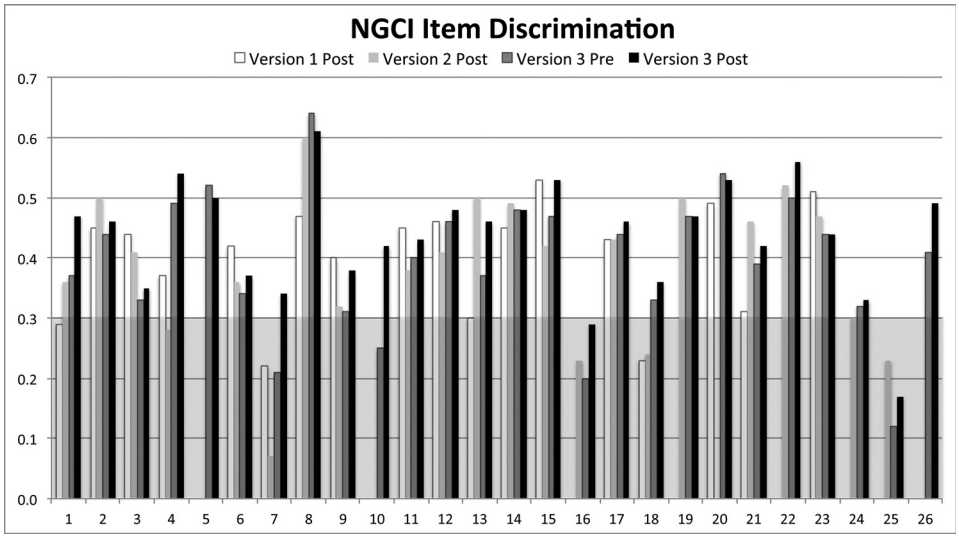

Figure 2. Version 3 NGCI item discrimination values, with accompanying values for Versions 1 and 2 where appropriate. The gray region represents values that typically indicate an item is a poor discriminator of student ability (less than 0.30).

(1994); Slotta (2011) suggest that novices can more easily construct a fundamental understanding of gravity when it is expressed as a Process, rather than as Matter. Table 4 below shows an example of how a question on the NGCI Version 1 (originally adapted from an open-response question in Williamson and Willoughby (2012)) was adapted to fit within the Process ontological category for Version 2. This change completely altered the meaning of the question, since the Version 1 wording of "surface gravity" was meant to refer to the gravitational field at the surface with choice "d" as correct answer, and the Version 2 wording clearly specifies an "interaction" with the orbiting body with choice "a" as the correct answer. Additionally, the reasoning statement provided within each of the choices of Version 1's question was eliminated because it may not represent all the possible ideas and reasoning that students use when choosing that particular choice. Once this ontological shift was implemented, only minor changes were implemented from Version 2 to Version 3. For example, the planets in the diagram were given shading and the mass labels were eliminated, and the distractor choices were modified based on student interviews.

5.2 Directionality Domain

Three items (Item 5, 10, and 26), shown in Table 5, were piloted for the first time in Version 3, all of which were intended to probe the Directionality Domain. Item 5 asks about the direction of the gravitational force for a ball following a parabolic trajectory, and, with a difficulty value of 0.28, it is the least difficult direction item on the NGCI. Item 26 was adopted from Dostal's Master Thesis (2005), in which it was used to probe physics students' understanding of gravity. These two questions probe student misconceptions related to the idea that the direction of the gravitational force is related to the direction of motion. Indeed, the most chosen distractor in each question is the arrow that points in the direction in which the object will be located a short time later (choice "c" for both items). Interestingly, this distractor is much more effective for Item 26. We believe this to be a subtle manifestation of the Boundary Model misconception, in which the gravitational force experienced by objects orbiting Earth is very different from that experienced by objects on its surface due to an atmospheric "boundary" or "threshold." We see this idea manifest in student responses to items within the Threshold Domain, such as Item 6 discussed in Section 5.5.

Item 10 requires students to reason about the combined effect, or superposition, of two gravitational forces. With a difficulty of 0.79, this question is quite complex because it also requires proficiency in the Force Law Domain via proportional reasoning with both mass and distance. Student interviews indicate that the most effective distractor, chosen more often than the correct choice, represents the idea that proportional changes to mass and distance are weighted equally in determining force. Interviews also suggest that the remaining distractors are chosen by students who reason with intuition rather than proportional reasoning. These naïve intuitions include the idea that objects can experience gravitational force from only one other object at a time, or simply that objects are pulled toward either the closer or the more massive object. Figure 2 shows that, despite its difficulty, this question is actually a very good discriminator of student understanding.

The addition of these three questions to Version 3 represents an effort to probe a wider range of student ideas in the Directionality Domain. Student ability to reason about direction in Versions 1 and 2 of the NGCI was mainly explored with questions about non-spherical objects. While these types of questions proved qualitatively valuable in the analysis described in Williamson and Willoughby (2012), introducing questions that use relative motion and superposition expanded the scope of the NGCI in measuring student understanding of how direction of gravitational force is determined.

5.3 Force Law Domain

Force Law questions on the final version of the NGCI test the common misapplications of the scientific model documented in Williamson and Willoughby (2012) while also gauging students' understanding of the proportionality of both mass and distance in determining gravitational force. For example, Table 6 shows how Item 4 on the final version of the NGCI probes student understanding of the role of distance by asking students to implement proportional reasoning when thinking about a changing Earth-Moon distance. Student interviews indicate that the most effective distractor is that which represents the misconception that distance, rather than distance squared, is inversely related to gravitational force. Table 6 also shows how the development of this question was informed by earlier versions, which, as evidenced by Figures 1 and 2, led to an item of appropriate difficulty and good discriminatory power. One can see that Version 2 of the question used a proportionality factor of two, and Version 1 used a factor of ten. In Version 1, the factor of ten was overly large. It was originally

Table 4. The evolution of Item 3 is an example of the ontological shift from Version 1 to Version 2 of the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are post-instruction

Version 1 $D = 0.49$, $r_{pb} = 0.44$	Version 2 $D = 0.45$, $r_{pb} = 0.41$	Version 3 $D = 0.37$, $r_{pb} = 0.35$
In Figure 3, planets A, B and C have the same mass, but A and C each has a heavy moon orbiting it, and B has a light-weight satellite orbiting it. Which planet has stronger surface gravity? a. (18.3%) Planet A, because it is able to hold a heavy object close to it. b. (18.3%) Planet A, because it is able to hold a heavy object close to it. c. (2.6%) Planet B, because only lightweight objects can orbit it. d. (22.2%) Planet C, because it is able to hold a heavy object that is far away. e. (51.3%) All have the same gravity, because they are all the same mass.	In Figure 4, planets A, B and C are identical. A and C each have a moon orbiting them, while B has an artificial satellite orbiting it, as shown in the diagram. *Each moon is twice the mass of the satellite*. Which planet has the strongest gravitational interaction with its orbiting body? a. (54.3%) Planet A b. (15.0%) Planet B c. (14.6%) Planet C d. (9.0%) All the same. e. (7.1%) Not enough information given to determine.	In Figure 5, planets A, B and C are identical. A and C each have a moon orbiting them, while B has an artificial satellite orbiting it, as shown in the diagram. *Each moon is twice the mass of the satellite*. Which planet has the strongest gravitational interaction with its orbiting body? a. (62.7%) Planet A b. (8.2%) Planet B c. (8.7%) Planet C d. (13.2%) Both Planets A and B e. (7.1%) All the same

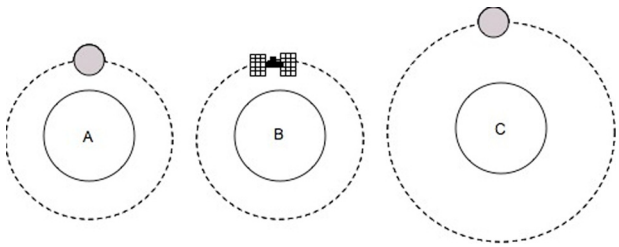

Figure 3.

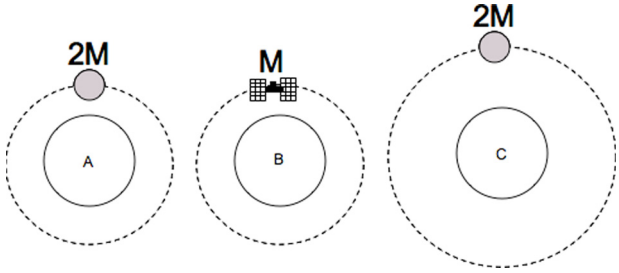

Figure 4.

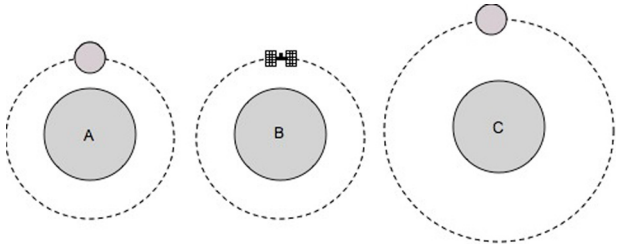

Figure 5.

Table 5. Three directional items introduced in Version 3. Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are post-instruction

Item 5 $D = 0.28$, $r_{pb} = 0.50$	**Item 10** $D = 0.79$, $r_{pb} = 0.42$	**Item 26** $D = 0.33$, $r_{pb} = 0.49$
A baseball is thrown at an angle so that it follows the dotted path. At the position shown, what is the direction of the gravitational force on the ball? See Figure 6. a. (5.1%) A b. <u>(71.6%) B</u> c. (13.7%) C d. (9.4%) D	Three planets are arranged as shown in the diagram. Planets X and Y each have mass *m* and Planet Z has mass *2m*. Planet X is a distance *d* away from Planet Y and a distance *2d* away from Planet Z. Which arrow (A-E) best represents the direction of the *total* (net) gravitational force on Planet X? See Figure 7. a. (13.2%) A b. (23.4%) B c. (36.3%) C d. <u>(20.6%) D</u> e. (6.1%) E	An Earth-orbiting satellite is shown at right. Which arrow, if any, best represents the direction of the gravitational force that it experiences? See Figure 8. a. (2.0%) A b. (6.1%) B c. (19.7%) C d. <u>(67.3%) D</u> e. (3.4%) E

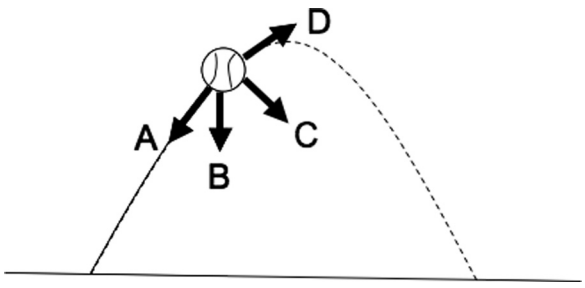

Figure 6.

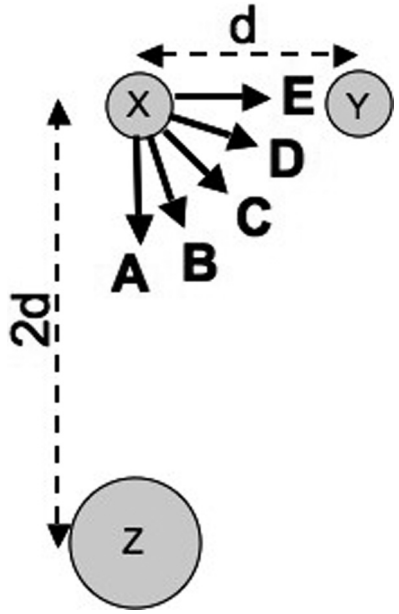

Figure 7.

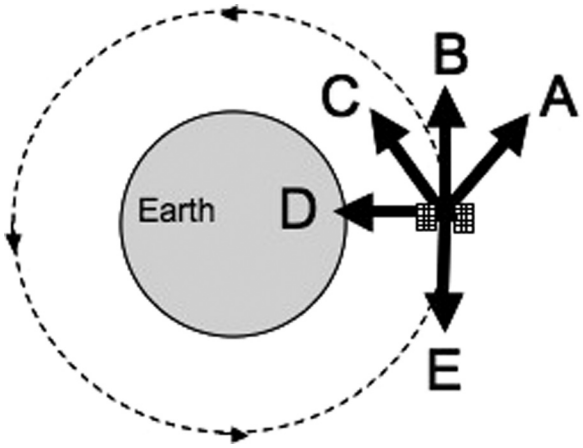

Figure 8.

chosen so that a student would simply need to remember that the gravitational force diminishes *quickly* or *drastically* to get the correct answer, but this proved to be too vague. In Version 2, distractor choices overly emphasized the factor of two. Interviews indicated that students did not consider the correct choice, "d. None of the above is correct," when presented with such effective distractors, making the question overly difficult and a poor discriminator. Additionally, one can see from Table 6 that choices that did not represent "weaker" were not effective distractors.

The conceptual change from the general "large" or "small" increase/decrease relationship in Version 1 to the proportional reasoning in Versions 2 and 3 was motivated by a greater awareness of Astro 101 instructors' practices. For example, several instructors use interactive engagement instructional strategies such as *Astronomy Lecture Tutorials* (Prather *et al*. 2013) and Ranking Tasks (Hudgins *et al*. 2006), which emphasize more rigorous inverse proportionality (i.e., if the distance increases by a factor of three the force decreases by a factor greater than three), and we wondered how frequently Astro 101 instructors require their students to do this. We emailed AstroLrnr listserv participants and 60 Physics and Astronomy Departments from around the United States. Out of 61 total respondents, 60 (98%) said that they at least show their students how proportionality works in the gravitational force equation. Of these 60, 54 (89% of the total respondents) also require that students demonstrate this ability (either in class or on homework or tests). Given the frequency with which this more rigorous inverse proportional reasoning is expected of the focus population, we deemed it appropriate to add this task to items on the NGCI. This represents a significant conceptual shift in the Force Law Concept Domain from Version 1 to Version 2.

Table 6. The evolution of Item 4 on the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are post-instruction

Version 1 $D = 0.39$, $r_{pb} = 0.37$	Version 2 $D = 0.87$, $r_{pb} = 0.28$	Version 3 $D = 0.60$, $r_{pb} = 0.54$
How would the gravitational force between the Earth and Moon change if the Moon moved out to *ten times* the distance of its present location?	How would the gravitational force between the Earth and Moon change if the Moon were located *twice* as far from Earth as it is now?	If the Moon were located *four times farther* from Earth as it is now, the gravitational force between Earth and the Moon would become…
a. (7.4%) Would be unchanged b. (17.0%) Would decrease only slightly c. (14.8%) Would be half as strong d. (60.9%) Would decrease a huge amount.	a. (72.3%) The force would become two times weaker. b. (9.0%) The force would not change. c. (4.9%) The force would become two times stronger. d. (13.1%) None of the above is correct.	a. (7.1%) Two times weaker. b. (35.7%) Four times weaker. c. (16.8%) Eight times weaker. d. (40.1%) Sixteen times weaker.

The most difficult item on the NGCI, Item 25, also falls within the Force Law Domain. As seen in Figure 1, this question maintained a high CTT difficulty value throughout the development process, and Table 7 shows its evolution through drastic changes. Item 25 probes how changes in mass and distance affect gravitational force by using a hypothetical comparison of two planets of different sizes but equal gravitational force for a person on their surfaces. At first encounter, Item 25 looks very similar to two other items on the NGCI (Items 14 and 17), where simple reasoning about mass, rather than size (or distance), can lead to the correct answer. Indeed, the most effective distractor corresponds to the idea that "same gravity implies same mass," which is only true if distance is the same. However, here, students must apply their reasoning about mass coherently with the idea of size (or distance), since a person *on the surfaces* of the planets would be at different distances from the centers of mass. So, even students that perform very well on the NGCI may get this question wrong because they do not double check the factor of distance. This likely explains the high difficulty and poor discrimination index of this question. However, even though CTT statistics did not improve throughout the development process, student interviews suggest that the changes did lead to a conceptually sound question that minimizes non-gravity related errors in student reasoning. Specifically, because Versions 1 and 2 compared Earth and Saturn, rather than hypothetical planets, students reasoned with their preconceptions about Saturn rather than their understanding of gravity. Some students were confused because it would be impossible for a person to stand on the surface of Saturn, and some students simply did not believe what the question was asserting. Furthermore, changing to a hypothetical scenario allowed distractor choices to represent common misconceptions in the Mixing of Forces Model. The purpose of this question is well aligned with the concept domains presented in Section 3, and it probes the vitally important student ability to correctly distinguish situations where either mass or distance is different. Therefore, despite its poor performance in terms of CTT statistics, we support the inclusion of Item 25 on the NGCI.

5.4 Independence of Other Forces Domain

One of the most prevalent and well-documented student misconceptions concerning the force of gravity is that it is confounded by other forces that students associate with magnetism, rotation, and air pressure. To probe this alternative model, we focused the Independence of Other Forces Domain on how the gravitational force relates to these specific physical phenomena. Because the multiple-choice questions in the Independence of Other Forces Domain were constructed with robust information from student-supplied descriptions of the effects of other forces presented in Williamson and Willoughby (2012), questions in this concept domain functioned as expected and changes tended to be minor. Table 8 shows one example of the evolution of a question, Item 8, which represents the types of changes that improved item performance within the Independence of Other Forces Concept Domain. Note the ontological shift in the phrasing of the stem from Version 1 to Version 2.

Table 7. The evolution of Item 25 on the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are post-instruction

Version 1 $D = 0.65$, $r_{pb} = -0.01$	Version 2 $D = 0.74$, $r_{pb} = 0.23$	Version 3 $D = 0.87$, $r_{pb} = 0.17$
The reason why the surface gravity on Saturn is similar to that on Earth is that: a. (55.2%) Saturn has the same mass as the Earth. b. (6.5%) Saturn has the same radius as the Earth. c. (34.8%) Although Saturn is much more massive, objects on the surface are farther from its center. d. (3.5%) All planets in the solar system have the same surface gravity.	The reason why your weight at the cloudy surface of Saturn is similar to that on Earth is that: a. (23.2%) Saturn has the same mass as Earth, and your weight depends only on the planet's mass. b. (9.0%) Saturn's large size would cause you to weigh more, but the rings counter the effect. c. (26.2%) Saturn is much more massive, but on its surface you are further from its center. d. (23.6%) Saturn is much more massive, but it is also much farther from the Sun. e. (18.0%) None of the above is correct.	For your weight to be the same on both Planets A and B… (See Figure 9.) a. (15.5%) Planet A must have a denser atmosphere because it is smaller. b. (63.3%) Planets A must have the same mass as Planet B. c. (7.13%) Planet B must be more massive and rotate faster. d. (12.5%) Planet B must be more massive because your location is farther from its center.

Figure 9.

Furthermore, by removing answer choices that involve density and size from Version 1 we created a question that more accurately probes students' use of the Mixing of Forces Model and can be answered with a standard five choice bubble sheet. As a result of student interviews, a choice for, "More than one of these," was added for Version 3 and became the most effective distractor.

5.5 Threshold Domain

The items of the NGCI that are aligned with the Threshold Concept Domain allow us to probe a combination of students' alternative mental models that were identified in Williamson and Willoughby (2012) related to certain limiting cases, such as low masses and large distances, as well as perceived "check points," such as a stable orbit (associated with the Orbital Indicator Model) or the presence of an atmosphere (associated with the Boundary Model). Below we discuss the evolution of two items (Items 21 and 6) that illustrate how students' ideas about these thresholds can be inextricably linked to observed motion.

Item 21, shown in Table 9, requires students to reason about the force between Earth and a list of specific objects. While the original open-response question and Version 1 appear to address the mass threshold idea directly, student interviews suggested that the final version elicits much clearer reasoning arguments from students. For example, a student who chooses "c" might reason that only objects whose motion is obviously affected by Earth experience a gravitational force with Earth (i.e., a person is stuck to the surface of Earth and the Moon orbits Earth). A student who chooses "d" might additionally conclude that even though Earth does not cause the Sun to move, Earth still exerts a gravitational force on the Sun because Earth orbits the Sun. While Haladyna, Downing, and Rodriguez (2002) do not recommend the final Complex Multiple-Choice format, this question is only moderately difficult for students and a good discriminator of ability.

The apparent weightlessness of astronauts is another observation that students use to reason about gravity. For example, 50.3% of the students in Williamson and Willoughby (2012) indicated that astronauts appear to float in

Table 8. The evolution of Item 8 on the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice postinstruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are postinstruction

Version 1 $D = 0.33$, $r_{pb} = 0.47$	Version 2 $D = 0.51$, $r_{pb} = 0.60$	Version 3 $D = 0.50$, $r_{pb} = 0.61$
Why does the Earth have gravity?	Why does the Earth exert a gravitational force on objects on its surface?	Why does Earth exert a gravitational force on objects on its surface?
a. <u>(67.4%) It has mass.</u>	a. (17.6%) It has an atmosphere.	a. (3.4%) It has an atmosphere.
b. (10.9%) It has a magnetic field.	b. (3.7%) It is very dense.	b. (5.3%) It has a magnetic field.
c. (5.7%) It is so large.	c. (16.5%) It has a magnetic field.	c. <u>(49.5%) It has mass.</u>
d. (6.5%) It is so dense.	d. <u>(48.3%) It has mass.</u>	d. (2.2%) It rotates.
e. (9.6%) It rotates.	e. (13.9%) It rotates.	e. (39.4%) More than one of these.
f. (0.0%) It has an atmosphere.		

Table 9. The evolution of Item 21 on the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are postinstruction

Open-response	Version 1 $D = 0.29$, $r_{pb} = 0.31$	Version 2 $D = 0.63$, $r_{pb} = 0.46$	Version 3 $D = 0.35$, $r_{pb} = 0.42$
How heavy or light does something have to be to create its own gravitational field? • 42.7% Every object has some gravitational field. • 18.90% Very, very heavy. • 7.0% Planet/moon sized. • 5.6% Only if it affects other objects. • 3.5% Not everything, needs to be tangible.	Approximately what minimum mass must an object have to gravitationally interact with another massive object? f. <u>(71.7%) Any mass will do.</u> g. (4.4%) About that of a person. h. (15.7%) About that of a moon. i. (8.3%) About that of a planet.	Of the objects listed at right, which experiences a gravitational force from the Earth? **Sun** **Moon** **Person** **Mars** a. (5.2%) None of them. b. (8.2%) Person only. c. (44.2%) Moon and Person. d. (5.6%) Moon, Person, and Sun e. <u>(36.7%) All of them.</u>	Which of these objects—**Sun, Moon, Person, Mars**—experience a gravitational force from Earth? a. (2.7%) Person only. b. (22.6) Moon and Person. c. (7.9%) Moon, Person, and Sun. d. <u>(65.1%) All of them.</u>

Table 10. The evolution of Item 6 on the Newtonian Gravity Concept Inventory (NGCI). Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are postinstruction

Open-response	Version 1 $D = 0.84$, $r_{pb} = 0.42$	Version 2 $D = 0.83$, $r_{pb} = 0.36$	Version 3 $D = 0.33$, $r_{pb} = 0.37$
Why do astronauts appear to float in their spacecraft?	Astronauts appear weightless in their spacecraft because…	Astronauts appear weightless in their Earth-orbiting spacecraft because:	An astronaut floating in her Earth-orbiting spacecraft…
• 50.3% There is no gravity in space. • 14.6% Gravity is much weaker in space. • 10.2% They are too far away from Earth or any massive body. • 9.7% They are in a constant state of freefall. • 2.1% The spacecraft's gravity isn't strong enough.	a. (16.1%) They are falling at the same rate as the spacecraft. b. (46.1%) There is very little gravity in space/outside earth's atmosphere. c. (14.8%) They are too far away from Earth or any massive body. d. (23.0%) There is no gravity inside the spacecraft.	a. (72.7%) They have escaped Earth's gravity. b. (3.4%) There is no air in the spacecraft. c. (17.2%) They are moving at the same speed as their spacecraft. d. (6.4%) The spacecraft's rocket engines counteract gravity.	a. (12.7%) experiences no gravitational force from Earth. b. (67.4%) still experiences a gravitational force from Earth. c. (19.4%) experiences a force from the spaceship that counters the gravitational force from Earth.

their spacecraft because there is no gravity in space. The results from previous studies (Ruggiero *et al*. 1985; Sharma 2004) and student interviews demonstrate that this misconception often is tied closely to the presence of an atmospheric boundary. However, even students who do understand that objects in space can gravitationally attract often do not readily have an explanation for why astronauts would appear weightless.

Item 6 on the NGCI, shown in Table 10, succinctly and efficiently probes the concept of floating as related to the concept of gravity in relation to an atmospheric threshold without relying on students to explain the effect in terms of orbital mechanics, which is not in the NGCI's intended scope. Early versions of Item 6 were overly difficult because they relied on non-gravity terminology. Specifically, interviews suggested that the connection between "apparent weightlessness" and "moving at the same rate as the spacecraft" in Version 2 was not readily accessible to students, as there are plenty examples of people on Earth moving at the same rate as a car or plane who do not experience weightlessness. Additionally, it was found that the word "falling" that was used in Version 1 generally means falling *down* for students, resulting in confusion and guessing. Not only is the final version more straightforward in terms of students' natural language, it more directly probes the idea of an atmospheric threshold related to gravity specifically. Again, the CTT statistics in Figures 1 and 2 show that these changes were effective in creating an item of appropriate difficulty and discrimination.

6. PHASE FOUR: VALIDATION

While the reliability of the NGCI is measured with the Cronbach's α statistic, the validity must be assessed via an argument-based approach (Kane 1992). Reliability refers to the consistency of an instrument's measurements, and validity refers to the interpretation one assigns to that measurement. In this section, we outline assumptions that we make about the interpretation of NGCI scores as a measurement of Astro 101 student understanding of Newtonian gravity and discuss evidence to support those assumptions.

Our first and primary assumption is that the NGCI measures understanding of Newtonian gravity. As discussed in Williamson and Willoughby (2012) and throughout this paper, we relied on a variety of sources to establish the most important aspects of Newtonian gravity for the target Astro 101 population and to integrate these into the design of questions. In Section 3, the Construction Phase, we outlined four concept domains of the NGCI. To gauge how well the items on the NGCI align with these concept domains, four astronomy education researchers independently categorized each item into one of the four concept domains. The results of this expert categorization of the items on the NGCI were evaluated by the Fleiss' Kappa inter-rater reliability statistic, which was calculated to be 0.80. According to Landis and Koch (1977), this represents substantial agreement. It is important to emphasize that the only disagreement was with which particular item falls within which particular domain; all experts agreed that all the NGCI items probe ideas within the concept domains as a whole. As further qualitative evidence, Section 5 shows how nine specific NGCI items exemplify these concept domains. Moreover, experts in Physics, Astronomy, and Education reviewed each version of the NGCI to ensure clarity, consistency, scientific accuracy, and adherence to best practices in instrument development. Twenty-four Physics faculty and graduate students evaluated the correctness of items by providing their answers to NGCI Version 3. Their average score was 96.7%. Quantitative analysis shows that student scores increased pre to post-instruction, and CTT item difficulty decreased from pre- to post-instruction while reliability increased, suggesting that students gained the ability to reason and answer questions correctly and consistently. Additionally, the wide range of observed normalized class gains (0.09–0.55) and effect sizes (0.13–2.12) indicates that the NGCI is sensitive to the effects of differences in Astro 101 instructional methods. These different lines of evidence indicate that higher scores are associated with greater understanding of gravity, lending validity to the idea that the NGCI measures understanding of Newtonian gravity.

Our second assumption asserts that students interpret the NGCI as intended. Specifically, students understand what each question is asking, and distractor choices function effectively as probes of student misconceptions. We have two primary pieces of evidence to support these claims. First, student-supplied responses to the open-ended questions in Williamson and Willoughby (2012) were foundational in describing students' naïve ideas. Examples of typical responses allowed question wording to mimic students' natural language. And, the most common misconceptions formed the basis for distractor choices. We can gauge the effectiveness of these distractors by looking at the distribution of student responses. Again, we provide concrete examples of this in Section 5. The second piece of evidence we have to support our assumption includes information from student interviews. These interviews were conducted on a voluntary basis with an incentive of $5 to participate. Eighteen students were interviewed while taking Version 2 of the NGCI, and seven were interviewed while taking Version 3. Students were given a copy of the NGCI, paper and pen, and they were asked to provide as much information as possible about their thought processes

while working through the survey, including whether they were confused, if they were wavering between answers, or if they were guessing. Students were audio recorded as they engaged in this "think-aloud" process (Willis 2005), while the interviewer engaged in "back channeling" by nodding and saying, "Okay," to encourage the students to continue (Bolton and Bronkhorst 1996), interrupting the students only to remind them to clarify or elaborate. The interviewer took notes during the interview and elaborated on relevant issues as soon as possible after the interview was over. As discussed in Section 5, these interviews were essential in determining how students interpreted questions and chose their answers. In some cases, information from interviews even suggested new distractors that proved to be popular choices. Interviews with students who were taking Version 3 of the NGCI became repetitive, indicating that students were interpreting and answering questions in a predictable way.

Taken together, the interpretive arguments in this section build a compelling framework that helps to establish the NGCI as a valid instrument in measuring Astro 101 student understanding of Newtonian gravity.

7. CONCLUSIONS

In this paper, we have made explicit the survey design and development process of the Newtonian Gravity Concept Inventory (NGCI). Our methods followed the best practices outlined in Benson and Clark (1982) and guided the organization of the paper. In Section 2, we outlined our approach to the Planning phase of instrument development, reviewing the results of the qualitative foundation developed in Williamson and Willoughby (2012). In Section 3, we described construction of four concept domains that included: (1) Directionality, (2) Force Law, (3) Independence of Other Forces, and (4) Threshold. We also discussed the process by which multiple-choice items were constructed. Section 4 described the test population and provided a Classical Test Theory statistical analysis of the reliability of the NGCI as well as measures of item difficulty and discrimination. Section 5 highlights the evolution of nine items to illustrate the iterative process by which we evaluated and modified the multiple-choice questions to ensure the conceptual breadth, scientific accuracy, and item clarity of the NGCI. Section 6 uses evidence from Williamson and Willoughby (2012) and Sections 3–5 to argue for the validity of the NGCI as a robust instrument for measuring Astro 101 student understanding of Newtonian gravity.

As a reliable and valid instrument for measuring Astro 101 student understanding of gravity, the NGCI can join the suite of Astro 101 concept inventories as a valuable instructional resource and research tool. To follow up on the results presented in this paper, we plan to use the NGCI to comment on instructional practices and student performance. In particular, we want to use the pre- and post-test data and demographic information from Tables 2 and 3 to explore NGCI gains and to tease apart the differences between students. We also hope to implement an Item Response Theory (IRT) analysis to expand our knowledge of NGCI item functionality and understanding of student ability along the construct of Newtonian gravity. Furthermore, we believe the NGCI is a useful instrument beyond the typical Astro 101 course, and we plan on measuring differences between college astronomy students and those in the typical introductory college physics courses. In the long-term, multi-institutional pre- and post-testing with the NGCI in many more classes can help to establish national norms on student understanding of Newtonian gravity. This data could also be used to analyze the effectiveness of traditional lecture compared to interactive engagement to build on the work of Prather *et al*. (2008) and to explore instructional strategies designed to focus on the known student naïve ideas and reasoning difficulties. This future research has the potential to significantly expand our knowledge of the preconceptions that college students bring to the classroom, and it can motivate a discussion of how Newtonian gravity should be taught so as to maximize learning.

ACKNOWLEDGMENTS

The authors would like to extend thanks to the many AstroLrnr participants and Physics and Astronomy experts who provided information about how they teach Newtonian gravity in their Astro 101 courses. The authors are especially grateful to the instructors Paul Robinson, John Feldmeier, Dan Loranz, and David Wittman who piloted the NGCI in their classrooms and collected some of the data for this paper. Additionally, the authors appreciate the thoughtful advice of Eric Brewe, whose suggestions helped to improve the quality of this paper.

NOTES

Note 1: Versions 1 and 2 were piloted locally at Montana State University post-instruction only. The average score for the 230 students who participated in Version 1 was 38.71% ($SD = 13.87$), and the average score for the 259 students who participated in Version 2 was 43.42% ($SD = 18.35$).

References

Ameh, C. 1987, "An Analysis of Teachers' and Their Students' Views of the Concept 'Gravity,'" *Research in Science Education*, 17(2), 212.

Asghar, A., and Libarkin, J. C. 2010, "Gravity, magnetism and 'down': Non-physics college students' conceptions of gravity," *Science Educator*, 19, 42.

Bailey, J. 2006, "Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties about the Properties and Formation of Stars," Ph.D dissertation, University of Arizona.

Bailey, J. M. 2009, "Concept Inventories for ASTRO 101," *The Physics Teacher*, 47, 439.

Bardar, E. M., Prather, E. E., Brecher, K., and Slater, T. F. 2007, "Development and Validation of the Light and Spectroscopy Concept Inventory," *Astronomy Education Review*, 5, 103.

Benson, J., and Clark, F. 1982, "A Guide for Instrument Development and Validation," *The American Journal of Occupational Therapy*, 36, 789.

Bolton, R. N., and Bronkhorst, T. M. 1996, "Questionnaire Pretesting: Computer-Assisted Coding of Concurrent Protocols," in N. Schwarz and S. Sudman (Eds.), *Answering Questions: Mehthodology for Determining Cognitive and Communicative Processes in Survey Research*, San Francisco: Jossey-Bass.

Chi, M. T. H., Slotta, J. D., and de Leeuw, N. 1994, "From Things to Processes: A Theory of Conceptual Change for Learning Science Concepts," *Learning and Instruction*, 4, 27.

Creswell, J. 2007, *Qualitative Inquiry and Research Design*, 2nd ed., Thousand Oaks, CA: Sage Publications, Inc.

Crocker, L., and Algina, J. 1986, *Introduction to Classical and Modern Test Theory*, Orlando, FL: Harcourt Brace Jovanovitch.

Ding, L., and Beichner, R. 2009, "Approaches to Data Analysis of Multiple-Choice Questions," *Physics Review Special Topics—Physics Education Research*, 5, 020103.

Dostal, J. 2005, *Student Concepts of Gravity*, Masters Thesis in Physics, Iowa State University.

Feeley, R. E. 2007. *Identifying Student Concepts of Gravity*, Masters Thesis in Science and Teaching, The University of Maine.

Gupta, A., Hammer, D., and Redish, E. F. 2010, "The Case for Dynamic Models of Learners' Ontologies in Physics," *The Journal of the Learning Sciences*, 19, 285.

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. 2002, "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education*, 15, 309.

Hambleton, R. K., and Jones, R. J. 1993, "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development," *Educational Measurement: Issues and Practice*, 12, 253.

Hornstein, S. D., Prather, E. E., English, T. R., Desch, S. M., and Keller, J. M. 2011, "Development and Testing of the Solar System Concept Inventory," *Bulletin of the American Astronomical Society*, 43.

Hudgins, D. W., Prather, E. E., Grayson, D. J., and Smits, D. P. 2006, "Effectiveness of Collaborative Ranking Tasks on Student Understanding of Key Astronomy Concepts," *Astronomy Education Review*, 5, 1.

Kane, M. T. 1992, "An Argument-Based Approach to Validity," *Psychological Bulletin*, 112, 527.

Kavanagh, C., and Sneider, C. 2006–2007a, "Learning about Gravity I. Free Fall: A Guide for Teachers and Curriculum Developers," *Astronomy Education Review*, 5, 21.

Kavanagh, C., and Sneider, C. 2006–2007b, "Learning about Gravity II. Trajectories and Orbits: A Guide for Teachers and Curriculum Developers," *Astronomy Education Review*, 5, 53.

Keller, J. 2006, "Development of a Concept Inventory Addressing Students' Beliefs and Reasoning Difficulties Regarding the Greenhouse Effect," Ph.D dissertation, University of Arizona.

Landis, J. R., and Koch, G. G. 1977, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33, 159.

Lindell, R. S., and Olsen, J. P. 2002, "Developing the Lunar Phases Concept Inventory," *Proceedings of the 2002 Physics Education Research Conference*.

Prather, E. E., Rudolph, A. L., Brissenden, G., and Schlingman, W. M. 2008, "A National Study Assessing the Teaching and Learning of Introductory Astronomy. Par I. The Effect of Interactive Instruction," *American Journal of Physics*, 77, 320.

Prather, E. E., Slater, T. F., Adams, J. P., and Brissenden, G. 2013, *Lecture-Tutorials for Introductory Astronomy*, 3rd ed., San Francisco, CA: Pearson Addison-Wesley.

Ruggiero, S., Don Minzoni, S. M., Cartelli, A., Alighieri, S. M. D., Dupre, F. and Vicentini-Missoni, M. 1985, "Weight, Gravity and Air Pressure: Mental Representations by Italian Middle School Pupils," *European Journal of Science Education*, 7, 181.

Schlingman, W. M., Prather, E. E., Wallace, C. S., Rudolph, A. L., and Brissenden, G. 2012, "A Classical Test Theory Analysis of the Light and Spectroscopy Concept Inventory National Data Set," *Astronomy Education Review*, 11, 010107.

Sharma, M. D., Millar, R. M., Smith, A., and Sefton, I. M. 2004, "Students' Understanding of Gravity in an Orbiting Space-ship," *Research in Science Education*, 34, 267.

Slotta, J. D. 2011, "In Defense of Chi's Ontological Incompatibility Hypothesis," *The Journal of the Learning Sciences*, 20, 151.

Smith, C. L., and Treagust, D. F., 1988, "Not understanding gravity limits students' comprehension of astronomy concepts," *The Australian Science Teachers' Journal*, 33, 21.

Wallace, C., and Bailey, J. M. 2010, "Do Concept Inventories Actually Measure Anything?," *Astronomy Education Review*, 9, 010116.

Williamson, K., and Willoughby, S. 2012, "Student Understanding of Gravity in Introductory College Astronomy," *Astronomy Education Review*, 11, 010105.

Willis, G. D. 2005, *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks, CA: SAGE Publications.

Wilson, M. 2005, *Constructing Measures: An Item Response Modeling Approach*, Mahwah, NJ: Lawrence Erlbaum Associates.