DEVELOPMENT AND CALIBRATION OF A CONCEPT INVENTORY TO MEASURE INTRODUCTORY COLLEGE ASTRONOMY AND PHYSICS STUDENTS' UNDERSTANDING OF NEWTONIAN GRAVITY

by

Kathryn Elizabeth Williamson

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Physics

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2013

APPROVAL

of a dissertation submitted by

Kathryn Elizabeth Williamson

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citation, bibliographic style, and consistency and is ready for submission to The Graduate School.

Dr. Shannon Willoughby

Approved for the Department of Physics

Dr. Yves Idzerda

Approved for The Graduate School

Dr. Ronald W. Larsen

STATEMENT OF PERMISSION TO USE

Kathryn Elizabeth Williamson

November 2013

ACKNOWLEDGEMENTS

Many people helped make this dissertation a success. First, I sincerely thank the physics and astronomy instructors who volunteered to pilot the NGCI in their classrooms, in addition to the thousands of students who agreed to share their ideas about gravity through interviews and written questionnaires. This data is the crux of the entire project.

Thank you to: Greg Francis and Shannon Willoughby for agreeing to take me on as a student at a critical time in my graduate career; Ed Prather for helping me to push deeper into the data and make connections that I might not have otherwise noticed; Janelle Bailey for kind and thorough feedback; and Colin Wallace and Carrie Myers for advice on statistical analysis. Colin was especially instrumental in helping me understand and use Item Response Theory.

I am also incredibly grateful for the Montana State University (MSU) Physics and Education Departments, the Montana Space Grant Consortium (MSGC), and the National Radio Astronomy Observatory (NRAO). These groups invested in my professional development in countless ways. I cannot thank them enough for providing a positive environment for my graduate education.

And, finally, I am especially thankful for my strong support system, including my parents, Robert and Jane, my brother, Michael, who is probably my biggest cheerleader, my husband, Adam Kobelski, for being my unwavering companion and confidante, and the rest of my truly good friends and colleagues, whose warmth continually reaffirms that this path I chose is where I belong.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF TABLES – CONTINUED

Table Page

LIST OF TABLES – CONTINUED

Table Page

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

ABSTRACT

The topic of Newtonian gravity offers a unique vantage point from which to investigate and encourage conceptual change because it is something with which everyone has daily experience, and because it is taught in two courses that reach a wide variety of students – introductory-level college astronomy ("Astro 101") and physics ("Phys 101"). Informed by the constructivist theory of learning, this study characterizes and measures Astro 101 and Phys 101 students' understanding of Newtonian gravity within four conceptual domains – Directionality, Force Law, Independence of Other Forces, and Threshold. A phenomenographic analysis of Astro 101 student-supplied responses to open-ended questions about gravity results in the characterization of students' alternative mental models and misapplications of the scientific model. These student difficulties inform the development of a multiple-choice assessment instrument, the Newtonian Gravity Concept Inventory (NGCI). Classical Test Theory (CTT) statistics, student interviews, and expert review show that the NGCI is a reliable and valid tool for assessing both Astro 101 and Phys 101 students' understanding of gravity.

Furthermore, the NGCI can provide extensive and robust information about differences between Astro 101 and Phys 101 students and curricula. Comparing and contrasting the Astro 101 and Phys 101 CTT values and student response patterns shows qualitative differences in each of the four conceptual domains. Additionally, performing an Item Response Theory (IRT) analysis of NGCI student response data calibrates item parameters for all Astro 101 and Phys 101 courses and provides Newtonian gravity ability estimates for each student. Physics students show significantly higher pre-instruction and post-instruction IRT abilities than astronomy students, but they show approximately equal gains. To investigate the differential effect of Astro 101 compared to Phys 101 curricula on students' overall post-instruction Newtonian gravity ability, linear regression models control for student characteristics and classroom dynamics. Results show that differences in post-instruction abilities are most influenced by students' pre-instruction abilities and the level of interactivity in the classroom, rather than the astronomy curriculum compared to the physics curriculum. These analyses show that the NGCI has broad capabilities.

CHAPTER ONE

INTRODUCTION

Over the last few decades, the disciplines of Physics Education Research (PER) and Astronomy Education Research (AER) have drawn on the growing body of research from cognitive science to inform pedagogy, research practices, and instructional tactics. The theory of Constructivism (Bodner 1986, von Glaserfeld 1981, Bransford, Brown, & Cocking 1999) is currently the most used and widely accepted framework for thinking about the learning process in physics and astronomy education research (McDermott 1991, Redish 2003), and as such it forms the foundation of this research project. This chapter provides the basics of Constructivism to motivate interpretations about cognitive structures, specifically describing “mental landscapes,” and methods of probing these cognitive structures, reflecting on the meaning of “assessment” and “measurement.” The chapter concludes with subsections about how and why these theories motivate the subject of the dissertation: developing and calibrating a Concept Inventory to measure introductory college astronomy and physics students’ understanding of Newtonian gravity. Chapter 2 reviews the historical progression of scholarly investigations of gravity to define what Newtonian gravity is, explore its implications, and provide a framework for interpreting student understanding of gravity. A literature review of studies on student understanding of gravity then explores how students reason about gravity in terms of its direction, as a force, in relation to other forces and factors, and in limiting cases.

Chapters 3 and 4 describe Phase I of the research project. Chapter 3 includes defining the scope and conceptual domain of the project and draws on the background literature to create accompanying open-ended, free-response questionnaires. Student-supplied responses to these questionnaires are coded and categorized, supporting the description of three mental models – The Boundary Model, the Mixing of Forces Model, and the Orbital Indicator Model – as well as other common misconceptions and misapplications of the scientific model. Chapter 4 draws on students' mental models and information from student interviews to create a multiple-choice assessment instrument, the Newtonian Gravity Concept Inventory (NGCI) that probes student understanding of Newtonian gravity within four conceptual domains – Directionality, Force Law, Independence of Other Forces, and Threshold. Classical Test Theory (CTT) statistics are used to motivate changes made throughout the iterative development process and to comment on item functionality. The reliability and validity of the NGCI is also discussed.

Phase II of the project is described in Chapters 5 and 6. Chapter 5 shows that the NGCI has applicability beyond the introductory college astronomy classroom, specifically in the introductory algebra-based physics population. A comparative analysis between astronomy and physics CTT item parameters and student response patterns is used to investigate differences between the astronomy and physics sample populations within each of the four conceptual domains probed by the NGCI. Chapter 6 then fits an Item Response Theory 2-Paramter Logistic Model to the composite astronomy and physics data, thereby calibrating NGCI item parameters independently of the sample population and measuring students' Newtonian gravity ability independently of the test items. The IRT student ability estimates are then used in a regression analysis to control for student demographics and classroom dynamics to investigate the differential impact of the physics curriculum compared to the astronomy curriculum on students' overall learning of Newtonian gravity. To conclude, Chapter 6 summarizes the project and discusses possible future classroom impacts and research directions.

Constructivism

In the mid-twentieth century, Jean Piaget revolutionized the theory of intellectual development in describing the process by which children draw on their immediate experiences and environments to learn about the world (Piaget 1952). Building on Piaget's theory, Constructivism posits that students do not arrive to the classroom as "blank slates," but rather, they bring robust mental landscapes. Students' prior experiences, knowledge, and beliefs, therefore influence what new information is perceived as relevant and how it is incorporated into their mental landscapes (McDermott 1991, Fosnot 1996). Piaget's "assimilation" describes the rather straightforward process by which new information is added to a learner's existing mental landscape, and "accommodation" refers to the more drastic and effortful restructuring that generally accompanies conceptual change (Vosniadou & Brewer 1987). Learning is enhanced when instructors are aware of the conceptions and ideas that students bring to the classroom. In particular, instruction that is tailored to explicitly confront preconceptions can provide the "cognitive dissonance" to motivate mental restructuring (Bransford, Brown, & Cocking 1999). Consequently, facilitating effective learning environments depends critically on characterizing the most common student conceptions and reasoning difficulties as well as reliably measuring how students' conceptions change as a result of instructional interventions.

Mental Models

Several useful frameworks for generalizing and describing learners' mental landscapes have been put forth. At the two ends of the spectrum are the ideas of "knowledge in pieces" and the theory of "mental models." Andrea diSessa (1993) espoused the more "atomistic" view of knowledge, in which some information is encoded in pieces for rapid use. These pieces of information need not refer to a robust explanatory "sense of mechanism," but rather reflect intuitive, phenomenological notions. These phenomenological primitives, or "p-prims," can take many forms but often lack systematic explanations. For example, a p-prim applied to the concept of Force might be, "more effort means more result" (diSessa 1993). Different contextual features of a problem may cue different p-prims (Chi, Feltovich, & Glaser 1981). Contrasting the atomistic perspective of knowledge is the more structured and organized perspective of "mental models" proposed by Vosniadou (1994). Mental models are "dynamic and generative" mental structures that are founded on certain ontological and epistemological presuppositions. Within this perspective, contextual features of a problem and new information lead to manipulation of the existing mental model structure.

In some sense, however, it is relatively unimportant which specific interpretation of mental landscapes is used (Viennot 1985). What is *observed* is that student reasoning is highly context dependent (Minstrell 1989, Bao & Redish 2006). Given two questions that probe the same concept, a student may answer one correctly and the other incorrectly. Therefore, a more generalized interpretation of a mental model as "a robust and coherent knowledge element or strongly associated set of knowledge elements" that is favored by Bao and Redish (2006) is adopted for this project. This broad interpretation includes simple and complex representations, correct and incorrect ideas, and a range of activating situations (Bao & Redish 2006). A misconception, then, is any aspect of the mental model that conflicts with expert interpretation. This more general interpretation is especially useful because the common reasoning patterns and understandings shared by students can be characterized through amassing response and interview data from multiple students and understood separately from how they manifest in any one case. Subsequently, a particular student' understanding of a particular topic in relation to context can be assessed empirically upon presentation of multiple conceptually similar questions via a probabilistic approach[1].

Assessment

The process of characterizing and measuring student understanding is known as educational assessment. As described by Kane (1992) via Mislevy (2006), "It [assessment] is reasoning from a handful of particular things students say, do, or make in a handful of particular circumstances, to more broadly construed inferences about what they have learned, have accomplished, or might do under various other circumstances." Assessment can be conducted in a variety of formats that depend on course goals, but in

---

[1] The specific probabilistic approach used in interpreting student understanding within this project, Item Response Theory, is discussed in Chapter 6.

all cases its foundation must include a qualitative argument to connect observations of students' actions to information about their mental landscapes and belief systems.

In an ideal world, the most effective assessment would include interviewing every student about her understanding of every course topic within every course goal and tailoring instruction on a one-on-one basis. However, this method is impractical at best, especially with large enrollment courses becoming the norm in higher education. Most instructors rely on homework, quizzes, and tests to assess student learning and to inform their teaching in the context of their course, but methods for discerning student performance in the larger context of national norms tend to be more ambiguous.

What is needed for such large-scale, comparative assessment is an efficient and reliable tool that can *measure* the a student's abilities in context with the larger population. Just as a meter stick measures length, an educational measurement tool must measure student understanding along a continuum of ability levels. Of course, measurement of a cognitive quality is subject to certain errors that measurement of a physical quality is not. Only when one has a trustworthy measurement device can one effectively research that which is being measured. Therefore, to measure the effects of instructional techniques and curricula on student learning, both PER and AER researchers have devoted considerable effort to creating reliable and valid measurement instruments.

Concept Inventories

Multiple-choice Concept Inventories (CIs) have become the measurement standard in the PER and AER communities for conducting large-scale assessment of student learning and the effects of instructional interventions. 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 correspond to common student naïve beliefs and reasoning difficulties, i.e. known misconceptions and alternative mental models. Each mental model is probed with multiple questions so as to tease out the effects of context dependence and minimize error (Bao 1999, Bailey 2006). Additionally, CIs focus on a narrow domain of topics that are central or foundational to the overall curriculum (Bardar et al 2007, Bailey 2009). Therefore, pre-instruction to post-instruction changes in students' scores on a CI can serve as a proxy for their learning due to instruction (Hake 1998, Prather et al 2009).

Furthermore, by investigating students gains on Concept Inventories, researchers have the power to evaluate the effectiveness of non-traditional techniques, curricula, workbooks, and texts (including *Physics By Inquiry* (McDermott et al 1996), *Tutorials in Introductory Physics* (McDermott et al 2002) and *Lecture-Tutorials for Introductory Astronomy* (Prather et al 2012)). For example, the Force Concept Inventory (FCI) and the Light and Spectroscopy Concept Inventory (LSCI) have been used to show that innovative, interactive, student-centered educational techniques greatly enhance learning compared to the traditional lecture-based format in introductory college physics (Hake 1998) and astronomy (Prather et al 2009) courses, respectively.

In accord with this foundational research and the Constructive framework, effort has been made to create measurement instruments that probe student understanding in other important topics taught in typical introductory college physics and astronomy courses. Some example topics include electricity and magnetism (Maloney et al 2001) to assess student learning in physics, and lunar phases (Lindell & Olsen 2002), star properties (Bailey 2006, Bailey et al 2009, and Bailey et al 2011), the green house effect (Keller 2006), light and spectroscopy (Bardar et al 2007), and the solar system (Hornstein et al 2011) in astronomy. With these tools, researchers, curriculum developers, and instructors have gained greater awareness about common student difficulties and how to maximize student learning of these topics; however, much work still needs to be done.

Focus of the Project

In order to progress the scholarship of PER and AER in assessing student learning, this project contributes an additional reliable, valid, and robust instrument to the "toolbox" of educational research through development, use, and calibration of a new Concept Inventory, the Newtonian Gravity Concept Inventory (NGCI). The NGCI was developed to focus on the particular reasoning difficulties of Astro 101 students (Chapters 3 and 4), but it also applies to introductory algebra-based physics courses (Chapter 5) and is calibrated to measure student learning in both courses (Chapter 6).

Why Astronomy Courses?

The majority of this study focuses on measuring conceptual change within introductory college astronomy, hereafter "Astro 101," students by researching their ideas about gravity. As opposed to the mostly pre-professional track (such as for medicine, dentistry, architecture, or veterinary sciences) students who take introductory algebra-based college physics, students who take introductory college astronomy tend to be non-science majors (such as business or education). Therefore, most of the 250,000 students who take Astro 101 each year (Franknoi 2002) will become society's business leaders, politicians, and school-teachers (Prather, Rudolph & Brissenden 2009). This student body clearly represents an important target population for developing scientific literacy and appreciation. However, Chapters 5 and 6 show that it is possible and instructive to extend the applicability of the NGCI to compare and contrast the Astro 101 student understanding and learning of gravity with that of introductory algebra-based physics ("Phys 101") students, providing a robust calibration of the instrument.

Why Gravity?

Gravity is a concept with which everyone has daily experience, so it is an excellent starting point for investigating and reshaping misconceptions in astronomy (and the physical sciences in general). A background review of Astro 101 textbooks indicates that gravity is an important concept that is covered in a typical astronomy course. Additionally, interviews with Physics experts (audio recordings available through APPENDIX C) indicate that gravity offers a good opportunity to think scientifically, a stated objective of many Astronomy teachers (Slater et al 2001) and a useful lifelong skill (Partridge & Greenstein 2003, Hobson 2008).

While the concept of gravity may be covered in only one or two lectures, gravity is useful for teaching other astronomical concepts (Smith & Treagust 1988), so students may be expected to apply their understanding of gravity to a variety of concepts throughout the course. For example, an understanding of gravity can provide a stronger understanding of the causes of tides, Kepler's Laws and motions of bodies in space, differences between irregular-shaped asteroids and spherical-shaped planets, star formation, Galileo's heliocentric model of the Solar System, the detection of extra-solar planets, and the grand structure of galaxies and the larger universe. On a broad scale, an understanding of gravity adds perspective to our place in the cosmos, including the distances over which bodies in space interact and the possibilities of spaceflight and exploration.

CHAPTER TWO

BACKGROUND

This chapter frames the study on introductory-level college student understanding of gravity within both historical context and extant educational literature. Kavanagh and Sneider 2007a and 2007b offer the most accessible and comprehensive review of the history of the science of force and motion in relation to gravity and student learning of gravity, hence their work helps to guide the following discussions. First exploring how philosophers and scientists came to a universal understanding of gravity serves to define gravity as a concept. The mathematics of Newton's theory are also explored to demonstrate the broad applicability, power, and importance of teaching this theory to students. Furthermore, because students' ideas about gravity parallel that of early scholars in many ways, reviewing the historical progression of the formulation of consistent theories of forces and motion offers valuable insight into how student misconceptions related to gravity can arise and be dispelled (Viennot 1979, Halloun & Hestenes 1985a and 1985b, Bar & Zinn 1998, Kavanagh & Sneider 2007a and 2007b). The accompanying educational literature review describes results from studies on student understanding of gravity in four conceptual areas that map to those discussed throughout this study: the direction of gravity, gravity as a force, gravity's independence from other forces, and the universality of gravity in limiting cases.

Historical Progressions in Understanding Gravity

With its clear prevalence in daily life, even the earliest scholars tried to describe and develop consistent theories of gravity. From Aristotle to medieval scholars to Newton to Einstein, humans continue to endeavor to understand this universal force of nature.

Aristotelian and Medieval Perspective of Gravity

Early understandings of gravity, force, and motion were heavily mixed with religious beliefs about Earth's place as the center of the universe subject to divine interventions. Earth was believed to be the center of the universe and surrounded by invisible spheres holding the Sun, planets, and stars (Kavanagh & Sneider 2007a and 2007b). In the Aristotelian perspective, "gravity" is an inherent property of an object – a measure of its heaviness – rather than an external force that pulls on the object. Falling is simply the object seeking its natural resting place at the center of the universe, with heavy, solid objects falling down fast and lightweight air and fire rising up. Motion other than falling requires continuous action or "impetus" (Ogborn 1985, Whitelock 1991). In the Aristotelian perspective, this propelling force on Earth comes from the medium (such as air) surrounding the object. In the heavens, there is no gravity and the divine force keeps the celestial spheres in constant motion.

In the medieval perspective, impetus is a motive force transferred from an active agent to an object. Rather than being created by the medium, the impetus is mediated or dissipated by the medium, such as resistance from air or contact with a surface, explaining why objects tend to slow down (Halloun & Hestenes 1985b) enough to allow gravity to take over. It was also assumed that the medium of space imparted no resistance to the celestial spheres. Therefore, the motion of celestial spheres did not require perpetual force, but only an initial imparting of impetus when God created the universe. Exploring these ideas led scholars of the time to define average and instantaneous velocity and acceleration, which formed the foundation for modern quantitative theories of motion.

Newtonian Gravity

With more accurate record keeping and the advent of the telescope, scientists such as Galileo discovered that some objects do not orbit Earth and that space is not bounded by fixed spheres, but appears to extend indefinitely, leading to the idea that objects might be able to travel without a continual force (Kavanagh & Sneider 2007b). Galileo also observed balls of different masses rolling down slopes of varying angles and concluded that their behavior under the influence of gravity is that of a uniform acceleration independent of an object's mass. No explanation for this observed phenomenon existed until Newton. In his *Principia* (translated by Cohen & Whitman 1967), Newton codified his three laws that united the concepts of force, mass, and acceleration to explain the motion of objects. His also presented his empirically derived (Cohen & Smith 2002) law of universal gravitation,

$$F = G\frac{m_1 m_2}{r^2},$$

which, together with his second law, explained why gravitational acceleration is independent of mass. Newton's law posits that *any* two particles with mass experience a mutual force directed toward one another, the magnitude of which is determined by the product of the two masses, m_1 and m_2, and the inverse distance squared, r^2, multiplied by a scaling factor, the gravitational constant, G. This simple law united the motion of falling objects on Earth with the motion of bodies in space under a common conception.

With advances in mathematics, Newton's Law had sophisticated implications for predicting complicated motions of bodies in space. First, to extend the applicability of Newton's Law, a gravitational vector field $\vec{g}$ can be defined. This interpretation implies that a test particle of mass $m_1 = m$ at location $\vec{r}$ experiences a gravitational force upon encountering $\vec{g}(\vec{r})$ produced by some larger distribution of mass $m_2 = M$ (see Figure 1). Then, Newton's Law can be expressed as:

$$\vec{F}(\vec{r}) = -m\vec{g}(\vec{r}),$$

where the minus sign simply indicates that the force is always attractive.

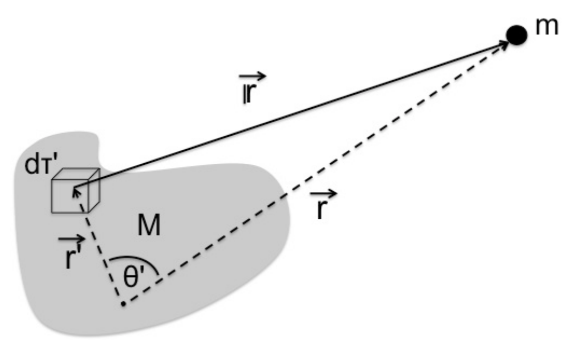


Figure 1: Definition of variables for derivations involving the gravitational vector field and scalar potential (Griffiths 1999).

The net gravitational vector field is the sum total of the contributions from infinitesimal mass elements of local density $\rho(\vec{r}')$ and volume $d\tau'$. By noting that $\vec{\mathbb{r}} = \vec{r} - \vec{r}'$, the gravitational vector field produced by M is then:

$$\vec{g}(\vec{r}) = -G \int \frac{\rho(\vec{r}')(\vec{\mathbb{r}})}{|\mathbb{r}|^3} d\tau'$$

Taking the divergence of both sides,

$$\vec{\nabla} \cdot \vec{g}(\vec{r}) = -G \int \rho(\vec{r}') \left(\vec{\nabla} \cdot \frac{(\vec{\mathbb{r}})}{|\vec{\mathbb{r}}|^3} \right) d\tau',$$

allows one to take advantage of the defining conditions of the Dirac delta function (Griffiths 1999), such that:

$$\vec{\nabla} \cdot \vec{g}(\vec{r}) = -G \int \rho(\vec{r}')(4\pi\delta(\vec{\mathbb{r}}))d\tau' = -4\pi G\rho(\vec{r}).$$

Furthermore, by integrating both sides over the Gaussian surface with volume V,

$$\int \vec{\nabla} \cdot \vec{g}(\vec{r})\, dV = -\int 4\pi G\rho(\vec{r})\, dV,$$

noting that $M_{enclosed} = \int \rho(\vec{r})\, dV$, and making use of the Divergence Theorem shows that:

$$\oint \vec{g}(\vec{r}) \cdot \overrightarrow{dA} = -4\pi G M_{enclosed}.$$

This result shows that the gravitational vector field at any location can be calculated with *only the mass enclosed* in the Gaussian surface through that location, greatly simplifying calculations of the gravitational vector fields of distributions with relatively simple geometries, such as spheres. For an extended, uniform, spherical distribution of mass M, the gravitational vector field reduces to:

$$g(r) = -G\frac{M}{r^2}$$

and the simpler version of Newton's Law can once again be applied:

$$F = G\frac{Mm}{r^2}.$$

However, for distributions of mass with no geometrical symmetry, the gravitational vector field must be approximated. This is accomplished through multipole expansion of the gravitational scalar potential. Because the gravitational force is conservative, the gravitational vector field can be expressed as the gradient of a scalar potential energy per unit mass (Fetter & Walecka 2003),

$$\vec{g}(\vec{r}) = -\vec{\nabla}V,$$

such that:

$$V = G\int\frac{\rho(\vec{r}')}{\mathbb{r}}d\tau'.$$

Using the law of cosines, it can be shown that:

$$\mathbb{r} = r\sqrt{1+\epsilon},$$

where

$$\epsilon = \left(\frac{r\prime}{r}\right)\left(\frac{r\prime}{r} - 2cos\theta'\right).$$

And, assuming $r' \ll r$, the scalar potential can be expressed as a sum of Legendre polynomials (Griffiths 1999):

$$V = G\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\int(r')^n\,P_n(cos\theta')\rho(\vec{r}')d\tau'.$$

Using any number of finite terms in this summation provides an approximation of the gravitational scalar potential, with more terms providing a more accurate estimate.

Therefore, in principle, for *any* known distribution of mass, the gravitational field can be calculated for *any* location in space, and, by extension, the gravitational force between *any* two objects can be known.

Pushing the mathematical capabilities of Newton's Law proved practically useful for understanding and exploring the structure of the cosmos. For example, it predicted the orbits of every planet and observable moon in the solar system with incredible accuracy, it predicted when comets would return, it led to the discovery of the planet Neptune in the mid-1800's, and by the mid-1900's it was used to launch spacecraft into planned trajectories through space. Despite its tremendous power, however, Newton's Law of Gravitation had some important conceptual and practical shortcomings of which even Newton was aware (Cohen & Smith 2002).

First, Newton's theory did not explain why an object's "gravitational mass," which mediates the strength of a gravitational interaction, is the same as its "inertial mass," which mediates an object's resistance to change in velocity. Additionally, Newton was extremely troubled by his theory's instantaneous "action at a distance," which required objects that were at exceedingly large cosmic distances from one another to immediately "know" of one another's mass and position. Furthermore, with advances in technology and observational techniques, it became apparent that Newton's theory did not account for the precession of the perihelion of Mercury. It was not until Einstein's theory of relativity revolutionized our understanding of the universe with the equivalence of motion subject to acceleration and gravity and the idea of a curved spacetime that the dilemmas of Newton's theory were resolved (Cohen & Smith 2002). However, in "ordinary" situations, Newton's theory is still frequently used and even preferred for its wide-ranging applicability and simplicity. For this reason, it is still Newton's theory that is taught in schools, and it is the tenants of Newton's theory that national educational standards expect students to understand (NGSS Lead States 2013).

Research Literature Review

The research on student understanding of gravity is quite extensive, and organizing this broad baseline of information is a formidable task. Kavanagh and Sneider (2007a and 2007b) found it instructive to separate the literature on student understanding of free fall from that of projectile and orbital motion. They further subdivided their publications into students' ideas and teachers' ideas, being careful to note that many cross-age studies show common difficulties for all groups. They also provided summarizing tables of the results of the studies they reference. In Williamson and Willoughby (2012), it was preferable to separate K-12 students' understanding of gravity from that of college students so as to distill what difficulties college students in particular might encounter when learning about gravity and to focus the methods of the research project. For this dissertation, however, the literature review is organized to mirror the four conceptual domains presented in Chapter 3. This provides the reader with a structure that clearly maps to discussions made throughout the dissertation. Furthermore, teachers' ideas about gravity are not emphasized in this review, so the interested reader is encouraged to refer to Kavanagh and Sneider (2007a and 2007b) and the references therein.

The Direction of Gravity

Young children's understanding of the direction of gravity depends heavily on their understanding of Earth's shape and its place in the universe (Agan & Sneider 2004). Nussbaum and Novak (1976) conducted one of the first studies that showed how children's personal experiences with falling objects moderated their learning about Earth's shape. They interviewed children with props and diagrams of a person dropping a ball from different latitudes on Earth. Their results suggested "Five Notions" of young children's ideas that progress in sophistication (see Figure 2): (1) Earth is flat, and any mention of a "round" Earth refers to curved sidewalks and hilltops, (2) Earth is round but objects can still fall down off of Earth to another surface, (3) Earth is round but objects can fall out into space, (4) Earth is round and objects fall to the surface but not necessarily toward the center of the planet, and (5) Earth is round, surrounded by space, with objects falling toward the center. These ideas were further researched with diagrams of people dropping a ball through a tunnel drilled through Earth in both a North-South and East-West direction by Mali and Howe (1979), Nussbaum (1979), and Sneider and Pulos (1983) to show that, in general, children progress from a strongly Earth-centered perspective of gravity to the more scientific perspective of Earth as a planet in the solar system as they age (Jones, Lynch, Resink 2007).

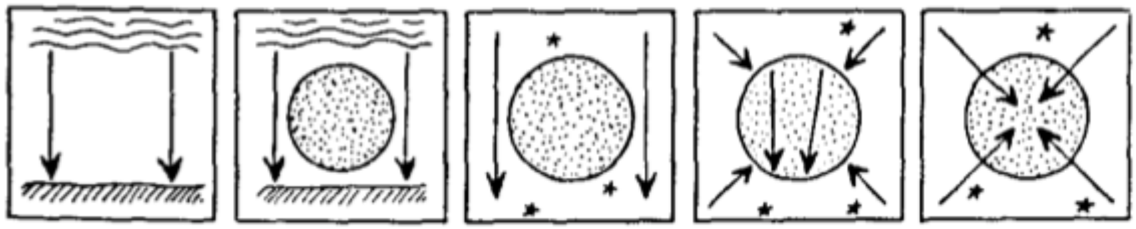


Figure 2: Nussbaum and Novak's (1976) characterization of children's understanding of Earth's shape and gravity.

Middle school, high school, and college students' understanding of the direction of the gravitational force has been researched by exploring student understanding of projectile motion and orbiting (Kavanagh & Sneider 2007a and 2007b). First, it appears that even as students age, some still hold the "straight down belief" observed in young children, reasoning that since gravity is straight down, motion must also be straight down. Caramazza, McCloskey, and Green (1983) asked college physics students about the motion of a ball on a pendulum after the cord had been snapped, McCloskey (1983) asked about an object pushed horizontally off a ledge, and McCloskey, Washburn, and Felch (1983) asked students where to drop a ball while walking so that it would hit a target. Findings from these studies suggest that many students expected the object to carry on with its previous motion for a small time (somewhat upward for the pendulum ball and forward for the ball pushed from the ledge) and then fall straight down. The studies' authors suggest that this misconception arises out of a strong association with the reference frame of the object.

On the other hand, some students reason that if an object is moving in a certain direction, there must be a force in that direction, gravitational or otherwise. Here is where one sees a hint of similarity to the medieval "impetus" theory. For example, Palmer

(2001) showed sixth and tenth grade Australian students diagrams of a ball (thrown up, at rest, and falling down), a person (jumping up from a trampoline, in a boat on water, and jumping down from a platform), and a brick (thrown up, buried under the ground, and falling down). Student responses suggested a variety of conceptions related to the direction of gravitational force:

(1) Gravity acts downward on falling objects.
(2) Gravity does not act upon objects that are moving vertically upwards.
(3) Gravity acts downward on stationary objects.
(4) Gravity does not act on stationary objects.
(5) Gravity acts upwards on objects that are moving vertically upwards.
(6) Gravity acts downwards on objects that are moving vertically upwards.
(7) Gravity does not act on falling objects.

Viennot (1979), Clement (1982), and Osborne (1983) showed that these misconceptions are exceptionally prevalent even after explicit instruction.

By asking students about the motion of objects in orbit around Earth or in hypothetical situations, researchers have gained insight into student conceptions about the direction of the gravitational force in more universal situations. For example, Dostal (2005) used force diagrams with multiple-choice and open-ended questions that required college algebra-based and calculus-based physics students to understand the principle of superposition in the context of gravity. The most common student difficulty when reasoning through a hypothetical diagram of two masses with a third placed between them corresponded to the idea that one object can block or stifle the gravitational interaction between two other objects. Furthermore, a majority of students believed that larger, more massive objects exert stronger gravitational forces on smaller, less massive objects than vice versa. Dostal (2005) also found that, while this misconception was reduced for students completing a tutorial worksheet designed to target Newton's Third Law in the context of gravity, it still remained strong even after instruction, with approximately a quarter of student responses corresponding to this misconception.

Gravity as a Force

As alluded to above, young children frequently view gravity as a property unique to Earth, describing gravity as the thing that holds people to the ground (Bar & Zinn 1998). Most high school and college students, however, seem to have an intuitive idea that mass and distance are important in determining the strength of gravity, but misconceptions arise in how this is applied. As discussed above and in the following subsection, many students reason with a minimum mass "threshold." Furthermore, one of the most well known student misconceptions about mass aligns with the Aristotelian idea that heavy things fall faster and the medieval idea of "impetus" (Champagne, Klopfer, & Anderson 1980, Gunstone and Watts 1985, Sequeira & Leite 1991, Graham & Berry 1993). This indicates that students do not have a working mental model for the differences between force and acceleration, which would allow them to understand how mass affects gravity. In regards to understanding the role of distance, students seem to understand that larger distance causes weaker force, but misconceptions arise in how quickly force diminishes with distance. Some students use a "farther means weaker" argument to explain why objects accelerate as they fall, reasoning that as objects get closer to Earth they speed up because they experience more force (Champagne, Klopfer, & Anderson 1980), and why objects in space appear weightless – they are too far to experience much of a gravitational force from Earth (Galili 1995).

Other research on student understanding of gravity as a force is in context with student understanding of forces in general, specifically in regards to Newton's three laws. Dostal's (2005) thesis represents the only study found that requires students to use mass and distance to quantitatively reason through Newton's Law of Gravitation. In particular, he found that many students do not use an inverse square law for distance, but rather just a one-to-one inverse relationship between force and distance, with this idea sometimes being more prevalent than the correct idea.

Gravity in Relation to Other Forces and Factors

Several studies (Watts 1982, Ruggiero et al 1985, and Driver et al 1994) show that some students' ideas parallel early scientific beliefs that gravity requires a medium through which to act, such as an atmosphere. Some students use this idea to explain that the strength of the gravity diminishes with altitude because the air is thinner and there is less air pressure (Gunstone & White 1980, Hestenes, Wells, & Swackhamer 1992, Feeley 2007). Furthermore, some students reason that the strength of gravity is zero in a vacuum such as space (Watts 1982, Sequeira & Leite 1991, Galili 1995). Bar and Zinn (1998) showed that some 9-18 year old students believe air to be a required medium for any "action at a distance" interaction, including magnetism, and furthermore, that gravity and magnetism are related in some way.

Other studies confirm that, indeed, in addition to air pressure, students confound gravity with magnetic and rotational forces, as well as other factors such as composition and distance from the Sun (Piburn 1988, Treagust & Smith 1989, Asghar & Libarkin 2010). For example, Piburn (1988) and Treagust and Smith (1989) showed college and tenth grade Australian students cards that showed a solar system in various contexts, such as with planets of different sizes, distances from the Sun, and rotation rates, asking which planet would be easiest for a rocket to "take off" from. They found that some students reasoned that the distance from the Sun affects the planet's gravity and that a planet's gravity depends on its temperature, composition, and orbital speed. While Dostal (2005) points out that these questions may be a bit confusing, since real rockets do need to consider the relative position of the Sun, it seems clear that students do hold deep misconceptions related to gravity and other factors. Asghar and Libarkin (2010) adapted diagrams from Nussbaum and Novak (1976) and Nussbaum (1979) and found that non-science major college geoscience students exhibit the same misconceptions as middle and high school students. Namely, they found evidence that the college students in their sample believed: that gravity exists on Earth because it rotates, that Earth's gravity is determined by its distance from the Sun and its place in the solar system, including in relation to the Moon, that gravity is the same thing as atmospheric pressure, and that Earth's gravity comes from its metallic core and is the same thing as magnetism. Only 12% of students in their sample held the correct scientific conception without simultaneously holding alternatives relating gravity to other forces or factors.

Gravity in Limiting Cases or Thresholds

The literature also suggests that students particularly struggle with the universality of gravity, believing that gravity acts selectively, or within some bounds or thresholds. This includes the idea that gravity only affects slow objects, inactive objects, or heavy objects (Kavanagh 2007a). For example, Watts (1982) found that some 12-17 year old

British students reasoned that a flashlight was affected more by gravity when turned on versus off. Bar et al (1994) also showed that some 9-18 year old students believe that objects can selectively choose or affect the gravity that they experience through strength or effort. For example, some students reasoned that a bird flying through the air can overcome the pull of gravity, clouds are strong and heavy so they do not fall, and a person in a boat on the water does not sink because he can paddle. Halloun and Hestenes (1985b) showed that even college physics students may reason that the effects of some forces depend on whether the active agent is living, finding that students reason differently about an object thrown from a person's hand versus dropped from a moving airplane. Finally, several studies use some variation of questions related to the behavior of a small object released or dropped by an astronaut standing on the Moon (Noce, Torosantucci, & Vincentini 1988, Ameh 1987, Reynoso, Fierro, & Torres 1993, Dostal 2005, and Feeley 2007). Results from these investigations show that many students believe that gravity requires atmosphere (discussed in the Gravity in Relation to Other Forces and Factors section above) or that gravity only acts on heavy things, i.e. that the object would float or drift away because there is some minimum mass threshold below which gravity stops pulling – what Feeley (2007) refers to as the "threshold model." Extending the Moon scenario, a common explanation for why astronauts were able to walk on the Moon is that they wore "heavy boots."

Studies that investigate student understanding of orbits offer particularly valuable insight about how students reason about the extent and reach of gravity. Many students exhibit the Aristotelian and medieval viewpoints that Earth is "special" and that objects on Earth behave differently than objects in space. The most prominent misconception among students is that this difference arises from or depends on the presence of the atmosphere. While some students confound air pressure with gravity, as discussed in the section above, many students believe that the edge of the atmosphere is an "upper limit" for the extent of gravity or a threshold that a spaceship needs to "break free" from (Watts 1982) and beyond which gravity no longer works (Feeley 2007). Sharma et al (2004) found that many introductory college physics students explained that an astronaut weighing himself on a scale in space would read zero on the scale because he is outside Earth's gravitational pull or that gravity in space in very weak. Galili (1995 and 2001) and Galili and Lehavi (2003) argue, however, that exploring student reasoning about gravity through the concept of weight may be flawed. They assert that the concept of "weightlessness" is confusing to students because physics courses generally equate "weight" with "gravitational force", but colloquial use of "weightlessness" generally refers to "apparent weight." To overcome this difficulty, they argue for adopting an operational definition of weight to be what a bathroom scale would read.

Discussion

Research on student understanding of gravity tends to parallel the historical progression but reflects more primitive conceptions. Similar to how the revolution from the Aristotelian and medieval to the Newtonian perspective required much evidence and convincing, an overarching theme in the research on student understanding of gravity is that students tenaciously hold to their preconceptions. Consistent with the Constructivist theory of learning (Fosnot 1996), students learn "commonsense theories" of force and motion by interacting with the world around them, which are usually taken for granted and unquestioned, making them particularly resistant to instructional interventions (Viennot 1979, Bliss, Ogborn, & Whitelock 1989, Halloun & Hestenes 1985a and 1985b, Kavanagh & Sneider 2007b). For example, Reynoso, Fierro, & Torres (1993) found that while older or more advanced students may use more sophisticated language to describe their reasoning, the underlying meaning of their words still shows strongly held misconceptions. Even students with strong academic and math backgrounds do not show an advantage in conceptual understanding of force and gravity (Champagne, Klopfer, & Anderson 1980). And, even when classroom experiments and demonstrations contradict with students' "commonsense" predictions, many students claim to see what fits their worldview and were particularly reluctant to accept the scientific view (Gunstone & White 1980 and 1981). In order to create learning situations that repeatedly cause the "cognitive dissonance" needed to encourage real and lasting conceptual change, instructors need to have a keen awareness of students' mental landscapes and reasoning difficulties.

CHAPTER THREE

CHARACTERIZING ASTRO 101 STUDENT UNDERSTANDING OF GRAVITY

While the literature on student understanding of gravity is wide-ranging, informative, and offers tremendous insight into common reasoning difficulties that students experience, only one of the studies reviewed (Feeley 2007) investigates the Astro 101 population in particular. What aspects of the misconceptions discussed above are most prevalent for this population of students? As discussed in Chapter 1 and argued throughout this thesis, reliably assessing and measuring students' understanding depends critically on characterizing their mental landscapes. This chapter serves to characterize the mental landscapes of Astro 101 students so as to inform the development of a multiple-choice Concept Inventory assessment instrument. It represents the foundational Planning phase of Benson and Clark's (1982) four-phase model for instrument development. Chapter 4 details how the results of the Planning phase were used to inform the three subsequent instrument development phases of Item Construction, Quantitative Evaluation, and Validation.

Method

The Planning phase of Benson and Clark's (1982) model calls for a clear narrowing of the topic domain being assessed and a broad investigation of the test population's understanding within that topic domain. To investigate Astro 101 students' understanding of gravity, this foundational process included generating a conceptual domain, administering open-ended questions, and categorizing and interpreting student-supplied responses.

Defining the Conceptual Domain

Because the topic of gravity has broad applicability and encompasses a variety of scientific ideas, it is necessary to define a conceptual focus that encompasses only the most relevant concepts of gravity that are emphasized in the typical Astro 101 course. To this end, during the Spring 2010 Semester a review of seven textbooks indicated that gravity is generally presented starting with the historical progression, followed by implications of falling, orbiting, and Newton's Laws in terms of planet formation, launching rockets, and tides. Additionally, general agreement among astronomy instructors and gravity experts was collected from eleven AstroLrnr email listserv members and through interviews with five Physicists/Astrophysicists[2] and one Geological scientist[3] (expert interview template in APPENDIX A and audio recording available through APPENDIX C). Discussions revealed that Newtonian gravity has many facets, with the most important aspects expressed in Table 1.

To assess student understanding of gravity with open-response questionnaires, these aspects of gravity were further narrowed to those in Table 2. First, the topic of tides was discarded from the conceptual focus. While a discussion of tides is prevalent in most astronomy courses, a thorough understanding requires higher levels of physics, such as the multipole expansion of the gravitational field discussed in Chapter 2.

---

[2] Greg Francis, Neil Cornish, Angela Des Jardins, Charles Kankelborg, and Sean Brannon
[3] James Schmidt

Table 1: The most important aspects of gravity.

The universality of gravity	The same physical principles of gravity apply to every object in the universe.
Influencing factors	The strength of gravity depends only on the mass of both objects and the distance between them (not size, density, rotation rate, composition, magnetism, etc.).
The scale and extent of gravity	A body's gravitational field falls off with distance but approaches zero asymptotically.
Tidal forces	The Moon's gravity causes the tidal bulge of oceans on the side closest to and opposite the position of the Moon.
Gravity is an interaction	Gravity is a force of attraction between two masses and is equal for both objects based on Newton's Third Law.
Equivalence Principle	Gravitational mass and inertial mass are the same quantity.
Action at a distance	Gravity is a non-contact force and cannot be blocked (for example, by another mass).

Second, Newton's Theory of Gravitation was chosen as the prominent focus over Einstein's theory. While Einstein's picture of gravity as the result of bending spacetime is often mentioned in introductory astronomy courses, it is rarely emphasized. And, at a level of sophistication beyond the scope of all Astro 101 courses, Einstein's theory predicts a coupling of gravity to rotation and magnetism, which could compound the known difficulties students experience when learning about gravity for the first time. Finally, the conceptual focus was narrowed to exclude Newton's three laws, as the Force Concept Inventory (FCI) is a more appropriate test for this domain, and, more importantly, it repeatedly shows that students have a poor understanding of forces (Hestenes, Wells, & Swackhamer 1992).

Table 2: The aspects of gravity probed with open-response questionnaires.

Direction	➢ Mutually attractive ➢ Toward center of mass ➢ Planet and Star formation ➢ Non-spherical objects
Force Law	➢ The role of both masses ➢ There is no minimum mass ➢ Inverse Square Law ➢ No boundary where gravity stops
Gravity unchanged by other factors	➢ Air pressure ➢ Magnetism ➢ Rotation ➢ Relative motion
Orbital Effects	➢ Why astronauts appear weightless ➢ Why planets do not fall into the Sun ➢ Why rockets can leave the Earth

As discussed in Chapter 4, additional insight was gained throughout the research process via analysis of student responses to multiple-choice questions[4] and student interviews[5], which helped to refine the conceptual focus shown in Table 2 into four, more distinct, conceptual domains– Directionality, Force Law, Independence of Other Forces, and Threshold – which informed the development of the multiple-choice Concept Inventory. The breadth of both correct and incorrect ideas probed within these final conceptual domains is expressed in Table 3. While it is important to keep in mind that all of these ideas are inherently related to those in Table 1 and Table 2, it is convenient and instructive to explore student understanding along these four particular domains. Specifically, within the context of this study, a student who understands Newtonian gravity can be defined as one who can correctly and coherently reason about: the direction of the gravitational force, the magnitude of the gravitational force, the

---

[4] APPENDICES I, L, and P

[5] See APPENDIX C audio recordings and APPENDICES D, H, N, R, and S for notes.

relationship between mass and gravitational force, the relationship between distance and gravitational force, how the gravitational force is independent from other forces, and the universality of the gravitational force in limiting cases.

Table 3: The final four domains that define the conceptual bounds probed by the multiple-choice Newtonian Gravity Concept Inventory. Ideas represent the range of student ideas uncovered throughout the research process, with those that are correct marked with an asterisk.

Directionality
Multiple Objects
• *Direction of the total (net) 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
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.

Table 3: Continued.

• 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

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

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.

Free-Response Questionnaires

In order to build on the extant literature and to get a basic sense of Astro 101 students' understanding of gravity during the Planning phase, a short questionnaire was created (APPENDIX B) and administered verbally during the Spring 2010 semester to 18 Astro 101 students via audio-recorded interviews (students S10-01 through S10-18 in APPENDIX C). Notes from these interviews are included in APPENDIX D. This helped to inform the development of three free-response questionnaires.

With a goal of focusing on the causes, effects, and extent of gravity, the free-response questions had little to no emphasis on equations. In order to reliably probe students' mental models, questions where students needed to have heard the answer before were avoided, and instead questions whose answers can be deduced easily by a true understanding of gravity were emphasized. These are generally different types of questions than those that are asked in textbooks at the end of the chapter. In total, 30 questions were developed for the questionnaires, and these questions were divided evenly into 3 versions: A, B, and C. The questions explore the strength of gravity in a variety of environments and contexts, such as on the Moon, in space, in free-fall, and on planets with orbiting bodies. During the Fall 2010 semester, approximately three weeks after a lesson about gravity, each section of introductory astronomy at Montana State University (MSU) was given thirty minutes to complete a different version of the questionnaire. One hundred forty-three students participated in Version A of the gravity survey, 137 students participated in Version B, and 32 students participated in Version C. Because responses to Version C were low in number[6], interpretation of student responses to this version should be viewed with caution. All three questionnaires and the results for each question (discussed below) are found in APPENDIX E, APPENDIX F, and APPENDIX G, respectively.

Follow-up, audio-recorded interviews were conducted (see APPENDIX C for recordings and APPENDIX H for notes) to establish credibility of the questionnaires and to gain insight into the complexities of student reasoning. As an incentive, students were offered the chance to win one of five $20 Target gift cards. Fifteen students volunteered for interviews (students F10-01 through F10-15). This included 5 students who completed version A, 9 students who completed version B, and 1 student who completed version C. All interviews were completed at the end of the semester, between four and seven weeks after the gravity questionnaires were given in class. Interviewees were asked to interpret their written responses and asked probing questions to bring up contradictions in their reasoning. Interviews typically lasted 20-30 minutes.

These free-response questionnaires and interviews represent a more foundational, wide-ranging assessment of Astro 101 students' ideas in order to serve three purposes: (1) to explore a wide range of student responses to different representations of gravity in a variety of situations, (2) to allow for analysis via a constant comparative method, as discussed below, and (3) to gain information for distracter responses for the multiple-choice Concept Inventory that represent typical student answers in their natural language.

---

[6] Version C was administered to a smaller section of students, the class met late in the afternoon, and the instructor introduced the questionnaire as purely optional.

Constant Comparative Approach

In analyzing student responses to the free-response questionnaires, the researchers (Williamson & Willoughby 2012) followed Creswell's (2007) model for a grounded theory constant comparative research approach, which has been used in many studies on student conceptual understanding (Bailey 2006, Sharma et al 2004, Asghar & Libarkin 2010). The central idea to constant comparative research is an open-coding format where the researcher collects data and allows categories to emerge naturally until the information gained is saturated. In constructing the emergent categories of student responses to each question, a phenomenographic approach was taken (Sharma et al 2004), which de-emphasizes correctness in favor of probing for a complete description of student ideas.

Student responses were compiled by typing responses into a single document and giving each response a number tag so that individual cases could be tracked. In an exploratory fashion, the student responses were read, re-read and rearranged to determine the emergent themes for each question. Once the general themes, or categories, were determined, a colleague (Willoughby) separately coded the numbered student responses into the determined categories. To illustrate how these codings were compiled and analyzed, consider an example of ten student responses (#1 - 10) to a question that has two conceptual categories, 'a' and 'b'. Table 4 shows how each rater in this example coded each student response. The raters agreed that four responses (N=4) fit in Category 'a' and three responses (N=3) fit in Category 'b'. The raters did not agree on the categorization of two responses (#5 and #9). Rater 1 coded response #5 in Category 'b' while Rater 2 coded it in Category ‘a’. Also note that both raters deemed response #2 inapplicable to the question.

Table 4: An example of the codings of student responses by two raters.

		Rater 1		
		Category ‘a’	Category ‘b’	Not applicable
Rater 2	Category ‘a’	1, 3, 6, 10	5	
	Category ‘b’	9	4, 7, 8	
	Not applicable			2

To measure the degree to which the two codings of student responses about gravity agreed, a matrix similar to Table 4 was used to calculate the Cohen’s κ inter-rater reliability statistic (Cohen 1960) for each question. Inter-rater reliabilities greater than 0.80 are generally accepted as good agreement (Landis & Koch 1977). If the categorizations for any question resulted in $\kappa<0.80$, discussions and clarifications were made in an iterative process until acceptable agreement was reached, but no explicit information was exchanged as to which responses each researcher placed in each category.

Results

APPENDICES E-G include a comprehensive summary of questionnaire results from student written responses. While the questionnaires contained a total of 30 questions, results from only 23 questions are presented. The other 7 questions were generally misunderstood by students or provided no new, previously undocumented, information. The inter-rater reliability statistic, κ, is shown at the end of each question. Each question is displayed with an accompanying three-column table. The first column of the table lists the conceptual categories; the second column lists the corresponding total number, N, of responses that fell within that category according to *both* raters. Note that the number N does not include student responses that were placed in different categories according to each rater, or incoherent or inapplicable student responses, as these responses were not useful in gauging students' understanding and reasoning on questions. Therefore, the sum of the N values across categories does not match the number of students who provided answers to each questionnaire. The third column in the tables of APPENDICES E-G expresses N for each category as a percentage of total respondents, i.e. N was divided by 143 for responses from Version A, 137 for responses from Version B, and 32 for responses from Version C.

APPENDIX H summarizes student interview results, including examples of possible working mental models. The excerpt from student F10-03 reasoning about question A1 during an interview below reveals how probing questions during interviews helped to elucidate relationships among misconceptions (audio-recording available through APPENDIX C)[7].

> Student (S): "Why do astronauts appear to float in their spacecraft?" My response is "there is no gravity in space."
> Interviewer (I): Okay, and do you still agree with that?
> S: Yes, I do.
> I: Okay, and what do you define as "space"?
> S: Anything outside of our gravitational pull.
> I: So is there a boundary where something is in our gravitational pull and then not?
> S: I think it gradually decreases from Earth's surface to nothing at some elevation above earth.

---

[7] Interview scripts are paraphrased for readability.

> I: Astronauts orbit Earth just outside our atmosphere, so I guess I'm asking if it [gravity] decreases fast enough to get to be zero just outside our atmosphere.
> S: Yes.
> I: So then is the gravity from the Earth zero at the Moon? Does the Moon feel Earth's gravity at its location?
> S: No, the Moon has its own gravity.
> I: But the Moon orbits the Earth, right?
> S: True.
> I: So *why* does it orbit the Earth?
> S: Because of the gravitational pull between the two masses.
> I: Okay, so it seems that you're saying that the Moon *does* feel the Earth's gravity.
> S: Yes.
> I: So why can the Moon feel Earth's gravity, but the space station can't?
> S: Because the space station isn't a big enough mass.

In answering that there is no gravity in space, the student may appear to conform to the common misconception that gravity needs air to act through. However, thinking about Earth's effect on the Moon helped this student to realize that Earth's atmosphere does not affect gravity. Therefore, a more appropriate description for this misconception is that gravity only applies to heavy objects, i.e. within a certain mass threshold.

Student Alternative Models of Gravity

The analysis of students' written responses and interview discussions allowed for a description of emergent thought patterns, which suggests potential student alternative "mental models" as described by diSessa (1993) and Redish (1994). In the subsections below, each emergent mental model is presented in the general sense as in Bao and Redish (2006) (see Chapter 1). In reading each subsection, it is important to keep in mind that each mental model is not independent – student misconceptions about gravity are generally inextricably linked. Also, these mental models are by no means a complete description of all of student reasoning about gravity.

The Boundary Model. Perhaps the most prominent emergent theme is a "Boundary Model," which re-conceptualizes the literature on student understanding of gravity across certain thresholds. A defining "boundary" could be the surface of a planet, the edge of the atmosphere, or even an orbit. It may represent a casing that encloses a planet's gravity, or a "check point" where gravity either disappears or diminishes. With the Boundary Model, one can explore several student responses. A combination of student responses to Questions A2, B3, and A1 illustrates the Boundary Model in very specific regions of Earth – inside Earth, above the surface, and beyond the atmosphere, respectively. For Question A2 many students (N=37, 25.9%) reason that the strength of gravity would not change if one traveled down a tunnel to the center of Earth. The most common explanation for this is that Earth's gravity is constant throughout, because Earth is one body (N=33, 23.1%). In this case it is Earth's surface that acts as a boundary, inside of which Earth's gravity is constant. However, Question B3 illustrates that most students (N=71, 51.9%) understand that Earth's gravity diminishes with distance from the surface. Considering farther distances from the surface of Earth, the majority of respondents to Question C1 (N=15, 46.9%) assert that one simply must leave Earth (and its atmosphere) to feel zero gravitational force from Earth. This is a well-document misconception (Watts 1982, Asghar & Libarkin 2010) and is supported by typical responses to Question A1 that assert astronauts appear to float in their spacecraft because there is no gravity in space (N=72, 50.3%). In the Boundary Model, there are many ways this could make sense: the atmosphere is the casing that holds gravity inside, or the extent of the atmosphere is simply an indicator of the presence of gravity. The latter explanation is clearly articulated in many student responses to Questions A1 and A3. For example, Interviewee F10-04 discussing his response to A1 asserts:

> S: Astronauts are able to float in their spacecraft because they've broken out of the atmosphere… once they've gotten past the point where Earth is able to sustain any kind of atmosphere, then that's where gravity is weak enough for it [floating] to happen. I don't think it's a correlation in any meaningful sense. It's not like cause and effect or anything.
> I: So it's not like the atmosphere is causing gravity?
> S: Where the atmosphere is, gravity is strong enough for there to be gravity. That's how I would think it.
> I: Atmosphere is more of an indicator of where gravity is working?
> S: Yeah, that would be how it works. Yeah.

Combining these typical responses, one arrives at several possible student models for discontinuities, or "boundaries," in Earth's gravitational influence. Variations of the Boundary Model in and around Earth are presented in Figure 3.

Also in accord with the literature, student responses indicate a prevalent misconception that there is no gravity in space. This surfaces in student responses to Question A1, but it is also prevalent in responses to Question B3, where 14.6% (N=20) of respondents reason that the strength of Earth's gravity on top of Mt. Everest is weaker than at sea level because Mt. Everest is closer to space. In Question B5, a few students (N=5, 3.6%) asserted that the Apollo spacecraft experienced no gravity because it was in space. And, in Question C10, 12.5% (N=4) of respondents asserted that two astronauts in space do not experience a gravitational force because there is not gravity in space.

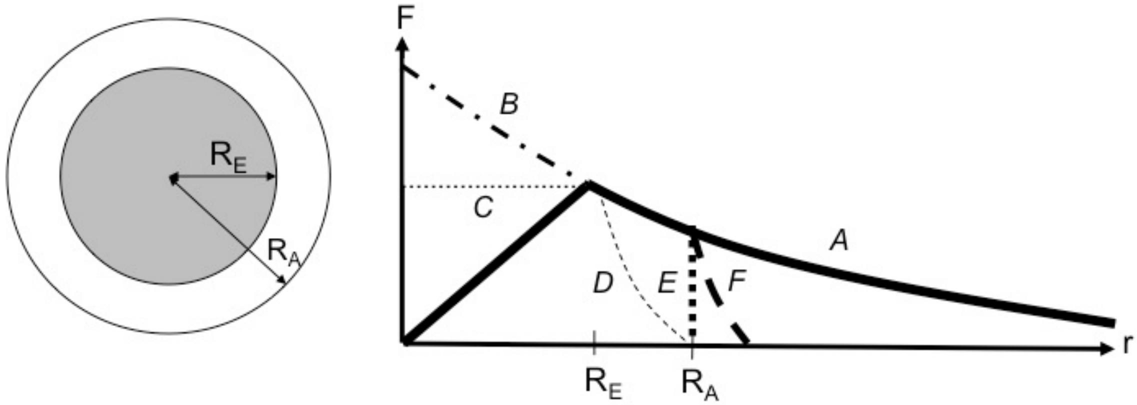


Figure 3: Variations of the Boundary Model for the strength of the gravitational force, F, that an object would experience as a function of distance from Earth's center, r. *(A) Solid Line:* The scientific model if uniform density is assumed, *(B) Dash-Dot Line:* the misconception that gravity is strongest at the center of Earth, *(C) Thin Dotted Line:* the misconception that Earth's gravity is constant throughout because it is a single mass, *(D) Thin Dashed Line:* the misconception that Earth's gravity drops to zero at the edge of the atmosphere, *(E) Thick Dotted Line:* the misconception that the atmosphere encases Earth's gravity, and *(F) Thick Dashed Line:* the misconception that gravity very quickly diminishes to zero outside the atmosphere. Students may hold a combination of these models.

Within the context of the Boundary Model, it may be reasonable to assume that any time a student asserts that there is no gravity in space, he is working from the idea that "space" is always outside the boundary, whether it be at the surface of a planet or past its atmosphere. As indicated in the interview transcript presented at the beginning of this Results section, the boundary may be different for different orbiting objects, extending out farther for more massive objects. The boundary may not define the edge of "space" at all, but rather the conditions for orbiting to occur, inside of which an object falls to a planet, and outside of which an object flies off into space. Responses to Question B1 use the Boundary Model in terms of orbits to explain why the planets do not fall into the Sun: "Gravity is pulling the planets towards then Sun but the gravity is not strong enough to actually pull them in," and, "They are so far away and the Sun doesn't have enough gravity," for example. In this case, the boundary divides regions where the Sun's gravity pulls objects in and where it simply holds them in orbit. While this may not be completely incorrect, as planets *are* in equilibrium, it is clear that students do not understand *why* the planets are in equilibrium.

The Orbital Indicator Model. An unexpected misconception that was uncovered is that the objects that orbit a planet can act as indicators of the strength of gravity on its surface. This was strikingly revealed by the majority of student responses to Question B8 (APPENDIX F). This question showed equally sized planets with orbiting satellites or moons at different orbital radii. Categories 'a' and 'b' show how the Orbital Indicator concept is generally applied in two opposing ways: (a) A planet's gravity must be stronger to pull heavy objects close to it, or (b) A planet's gravity must be stronger to reach out and hold heavy objects that are far away. A typical response for Category 'a' is as follows:

> "[Planet] A would have no gravity, B and C satellites would feel the same gravitational pull, D's satellite would experience less gravitational pull, E and F's moon would experience the same pull, G's moon would experience less."

This student claims that Planets B and C have *more* gravity than Planet D, and, similarly, Planets E and F have *more* gravity than Planet G. This implies that the student thinks that the gravity of planets B and C must not reach as far, and must therefore be weaker than the gravity of planet D. The reasoning here is that a planet's gravity must be stronger to pull an object closer to it.

A typical student response falling in Category ‘b’ of the Orbital Indicator Model is as follows:

> “[Planet] A: not much gravity, B and C [have the] same gravity, D [has the] most gravity because it can have a satellite [at the] furthest [distance]. E, F, and G have tides. G has the strongest gravity [because its] moon is furthest away, but E and F both have relatively strong gravity. F might not have tides b/c of [its] two moons.”

Like the previous student who fell in Category ‘a’, this student reasons that Planet A must have weak surface gravity because there is nothing orbiting it. However, unlike the previous student, this student claims that Planets B and C have *less* gravity than Planet D. And, because the student separated his predictions for the strength of gravity according to planets with artificial satellites and those with moons, he possibly expects Planets E and F to have *less* gravity than Planet G.

These two aspects of the Orbital Indictor Model come up again in Question C7 (APPENDIX G). Instead of satellites or moons orbiting a planet, different ring systems are depicted. Again, Categories ‘a’ and ‘b’ represent the two opposing applications of the Orbital Indicator Model. Category ‘b’ of Question C7 indicates that planets with less material in orbit have a stronger gravity. A student response in this category is: “The planet with more mass does not have rings, and [for planets with] less mass the planet carries the surrounding objects around it in orbit instead of having it become part of the planet.” Presumably the student is referring to “gravity” when he refers to “mass.” This implies that planets with strong gravity will cause all orbiting bodies to orbit close or even to spiral inward and impact the planet. If a planet’s gravity is weak, then it may allow some far away objects to steadily orbit. Again, this is analogous to Question B8

Category ‘a’ (APPENDIX F), where planets with stronger gravity are able to hold orbiting bodies at smaller radii.

Conversely, such as in Question B8 Category ‘b,’ students in Category ‘a’ of Question C21 predict that the gravity on a planet with more rings and thicker rings must be stronger to hold so much material around the planet. For example, one student asserts, “If E can keep such far-away rings in check, it is exerting more force than B is on its closer-in rings.” Again, this implies that a planet’s gravity must be strong to reach out and hold heavy objects that are far away.

The Mixing of Forces Model. Questionnaire results overwhelmingly support the well-documented (Piburn 1988, Smith & Treagust 1988, Chandler 1991, Graham & Berry 1993, Sharma et al 2004, Feeley 2007, Asghar & Libarkin 2010) misconception that gravity is confounded with magnetism, rotation, and atmospheric pressure. Question A3 explicitly asks students how gravity is related to atmosphere (if at all), Question A9 (APPENDIX E) explicitly asks how magnetism affects gravity (if at all), and Question B7 (APPENDIX F) explicitly asks how rotation affects gravity (if at all). Perhaps it is understandable for these questions that the majority of students assert that the atmosphere affects the strength of gravity and strong magnetic fields correlate with strong gravity. However, students readily incorporate magnetism, rotation, and atmospheric pressure into their responses to questions that do not address these forces explicitly. All three of these forces are popular responses in Question B2, which asks, “Besides gravity, are there any other forces that hold us to the surface of the Earth?” While it is encouraging that the majority (56.2%, N=77) of students understand that gravity is the only force holding objects on Earth, 12.4% (N=17) of students believe that atmospheric forces contribute, 6.6% (N=9) believe rotation and orbital dynamics contribute, and 2.2% (N=3) believe magnetic forces contribute. A few students who answered Question A6 even reasoned that people at different latitudes on Earth would experience different forces of gravity because of variations in the magnetic field (2.8%, N=4) or variations in rotational speed (4.2%, N=6). A few students refer to these forces again in other questions such as A2 and C2.

Responses to Question B1 (APPENDIX F), "Why don't the planets fall into the Sun?" illustrate this trend, with 5.1% (N=7) of students claiming that rotation of the planets and/or Sun counteracts the inward pull of gravity and another 5.1% (N=7) claiming that repelling magnetic forces counteract the attractive gravitational forces. Responses in category "g" of Question B1 indicate an interesting misconception that could potentially be tied to magnetism. Students in this category reason that planets do not fall into the Sun because, "The planets have their own gravity that repels them from being drug in," or, "They (the planets and the Sun) are pulling against each other." The comment, "Gravity is pushing away from the Sun," indicates that some student *may* believe that gravity is analogous to magnetism, at times attracting and at other times repelling.

A "Mixing of Forces" model is proposed to describe these types of student responses observed. Magnetism, rotation, and air pressure seem to be important factors in understanding gravity for some students. The "Mixing of Forces" model could also be used conjunction with the "Boundary" concept. To many students, these three factors can modify the reach, or extent, of gravity. Drawing on the literature review of Chapter 2, perhaps this can be tied to misconceptions about force and superposition. Perhaps students are drawing on everyday experiences where they (or objects they hold) "feel" lighter or heavier, and they attribute changes in apparent weight to changes in gravity.

Disparate Misconceptions. A few unexpected misconceptions surfaced from the student-supplied responses that do not seem to relate to a specific "mental model" of gravity – they appear to stand on their own as disparate topics. The most surprising misconception observed is the belief that South is "down." This has been documented in young children (Nussbaum & Novak 1976) but certainly not college students. However, 5.6% (N=8) of the student responses to Question A6 (APPENDIX E) indicate that a person at the North Pole of Earth experiences a stronger gravitational force than a person at the South Pole. Examples of student responses in this category include, "Person A is at the top and wouldn't fall off," and, "Person A is at a position where gravity is pulling them directly down." None of the students who fell in this category volunteered to be interviewed, which prevented further exploration of this misconception.

Another interesting misconception is the idea that gravity is a force that pulls an orbiting body along in its orbit, or that gravity establishes an orbit, which then "holds" the object. Responses in Category "c" of Question B1 (APPENDIX F) seem to indicate that this is how some students understand orbiting, reasoning that the planets do not fall into the Sun "because the orbit has been established," or, "because their orbits keep them in place." One student even explicitly stated, "When planets are in orbit, the force of gravity between the two bodies is not a direct pull towards each other." Perhaps this is loosely tied to the "Mixing of Forces" model, where the force of gravity is intertwined with rotation or centripetal forces, or perhaps it is just an extension of the "Orbital Indicator Model."

One more interesting misconception observed echoes the results of Dostal (2005): that gravity can be blocked or stifled, such as when the two objects are not in view of each other. This was most prevalent in responses in Category "c" of Question B5 (APPENDIX F). Students in this category reasoned that when the Apollo spacecraft was behind the Moon, "the Moon was in the way of Earth's gravity," or the Moon "interfered with" or "interrupted" Earth's gravity. One student stated "Earth had no effect on the spacecraft once behind the Moon. It would take effect again once it had view of Earth." This misconception could potentially relate to that of Category "b" in the same question, where some students appear to believe that an object can only experience the gravity of one object at a time. In this case, when the spacecraft is near the Moon, some students wrote, "the Moon's gravity was no longer competing with the Earth's," or, the spacecraft "was in the Moon's gravity, not experiencing the Earth's gravity at all." Again, this could potentially be tied to the "Mixing of Forces model," but it seems more related to a misunderstanding of superposition. Perhaps these students are too literal in their interpretation of gravity as a force of attraction between *two* objects.

<u>Student Misapplications of the Scientific Model</u>

As illustrated in the discussions in this section, students can have diverse mental landscapes of alternative models about gravity. However, this does not imply that they cannot also hold misconceptions related to the correct scientific model. The subsections below discuss three main ways in which student responses indicated a failure to use the scientific model correctly: misapplication of mass, misapplication of distance, and misunderstanding what causes or creates gravity.

Misapplication of Mass. Students misapply the factor of mass in determining the force of gravity in two ways: (1) assuming that only heavy objects can gravitationally interact, and (2) confusing mass and density. Responses to Question A4 show that 35% (N=50) of students believe that there is a non-zero minimum mass requirement for an object to gravitationally interact with other objects. Of these 35%, one tenth (N=5) of students come close to the scientific understanding by asserting that, for example, an object must be "probably a pound or more." An additional 20% (N=10) assert that it must be planet or moon sized. More ambiguously, 54% (N=27) simply state: "very, very heavy." This misconception surfaces again in Question C10, which shows two astronauts in space far away from other objects. Only 37.5% (N=12) of respondents correctly assert that the two astronauts experience a force of gravity from each other. However, 18.8% (N=6) of respondents say that the astronauts feel a force of gravity from distant objects. While this group of students appears to understand that the strength of gravity diminishes to zero only over infinite distances, they do not appear to understand that any mass can interact gravitationally with other objects. Perhaps this is a more general application of the "threshold model" presented by Feeley (2007), in which students reason that a pen released near the surface of the Moon would not be heavy enough to fall.

It appears that students have an intuitive understanding that more "stuff" means more gravity. This appears to be a phenomenological primitive, or "p-prim" (diSessa

1993), that students unquestioningly apply in their reasoning about gravity. It often serves them well, since more *mass* really does mean more gravity. However, many students assume that higher density always implies higher mass, without taking size into account. Because the density of a planet indirectly provides information about its mass (and hence its gravity), this misconception is particularly difficult to untangle. Alternatively, one must not discount the possibility that students are working from the correct mental model and simply using the wrong terminology, i.e. maybe students understand gravity, but they do not understand the difference between mass and density. Question A2 uniquely addresses this subtlety by asking students to imagine a thin tunnel that reaches to the center of Earth. Some students reason that gravity would increase because “the center of the Earth has a greater mass.” Of course, the center of Earth cannot contain more mass than the entirety of Earth, so it is reasonable to assume these students mean “density” when they say “mass.”

However, some students do seem to understand the difference between mass and density and show a bias toward the factor of density when reasoning about gravity. For example, a common alternative response to Question A2 is, “the densest matter of Earth is in the core.” Some students go so far as to consider density as the primary factor and mass as the secondary factor that determines the strength of gravity, as demonstrated in Interviewee F10-04’s response to Question A4, which asks how heavy something has to be to have gravity:

> “I think it’s gotta be pretty heavy… Something that’s incredibly massive and incredibly light wouldn’t necessarily be as much gravity as something that’s equally massive but far denser. So I think it has something to do with size, but more to do with density and mass and such, because if the Sun was far

> larger it wouldn't be able to attract planets to it. Like I learned that when the Sun becomes a red giant it's going to push all the planets out, and I figure that's why – because it's so much less dense as a red giant than something like a yellow star like it is right now."

This student is making a genuine effort to think scientifically by linking his reasoning with other information that he learned in class. He applies this same mental model throughout his interview responses. For Question A6, he reasons that a person at the equator would experience stronger gravity because Earth bulges, and, "assuming the Earth's density is constant," mass becomes the determining factor, and, "there is more stuff between the person and the center of the Earth." Even though this student has misconceptions about the relationship of density, mass, and gravity, he works from a consistent mental model.

Misapplication of Distance. Many students understand that distance between two objects is an important factor in gravity; however, misconceptions occur in how this is applied. For instance, many students become attached to the word "radius." Scientists often use the variable *r* for "distance," but students may interpret this too literally and deduce that *r* stands for "radius," and a smaller planet has stronger gravity because it has a smaller radius. This is exemplified in Category "h" of Question A8 (APPENDIX E), with a typical student response: "A is strongest, D is weakest. Masses will be the same but the radiuses [sic] change. The smaller the radius the greater the [gravitational] force." This misconception crops up again in Category "e" of Question B4 (APPENDIX F). One student reasoned that Pluto's gravity would be stronger than Earth's "because the surface is closer to the core."

Finally, an alternative misapplication of the distance variable is mistaking distance as the distance from the surface of a planet. A quote from Category "e" of Question A8 is an example of a student considering the distance from the surface of the planet: "The darker the planet the more dense. Gravity is affected by mass and distance. In the darker shaded planets they will experience more gravity. In the lighter shaded planets they are closer to the planet causing greater gravity."

Misunderstanding What Causes or Creates Gravity. Many students do not understand *where* gravity comes from. One interviewee explicitly states, "I don't know what creates gravity. I just know it exists." Despite understanding that mass is an important factor in determining gravity, many students fail to make the connection that the interaction of two masses is in fact the *cause* of the gravitational force. Because examples and problems in the classroom primarily involve spherical bodies, the center of the object is frequently referenced and students begin to think that the center is somehow different or more relevant than the rest of the mass. In fact, some students reason that the geometric center of an object is (1) where gravity originates, or (2) the location to which gravity pulls everything.

Question A2 (APPENDIX E) illustrates that many students believe the center of Earth to be somehow different gravitationally than the rest of Earth. Nearly a quarter (23.1%, N=33) of students reasoned that gravity would be stronger at the center of Earth than at the surface, with an additional 4.9% (N=7) implying that this is because the center is the source of gravity, providing comments such as, "we are closer to the source of the Earth's gravity, where gravity originates," and, "you are getting closer to the mass exerting the gravitational force." The former comment may provide insight into another facet of this misconception – it is not necessarily the center point where gravity originates, but a central region. Some students draw on their knowledge that the core of Earth is compositionally different than the crust. It is denser, and a student who believes that density, rather than mass, determines gravity would reasonably conclude the center exerts a stronger gravitational force than the crust. Also, the core of Earth is primarily iron, which could potentially factor into a student's mental model that includes magnetism. And finally, a student who believes pressure (atmospheric or otherwise) to be associated with gravity would logically conclude that the center of Earth exerts a stronger gravitational force than the rest of Earth.

Many responses to Question C22 (APPENDIX G) further demonstrate the misconception that gravity originates from the center of an object. Depicting a planet whose center of mass is offset from its geometric center unambiguously untangles student conceptions. While 34.4% (N=11) of respondents correctly drew an arrow representing the gravitational force toward the center of mass, 21.9% (N=7) drew an arrow to the geometric center. In the latter case, one student adds that, although gravity still points to the geometric center, "it may be weaker than normal." Again, students may understand that mass is a factor in a planet's gravity, but they do not understand its role exactly. Alternatively, a few students (9.4%, N=3) appear to believe gravity to be caused by a *lack of mass*, as these students drew arrows pointing into the hole of the planet. Since version C had so few respondents, one can only speculate as to students' reasoning in this process. However, arrows to the hole may allude to the misconception that gravity is measured by its effects on other bodies – a person could fall into the hole, but would merely remain on the surface at the side without a hole.

Student responses to Question A2 (APPENDIX E) unveil the subtle interplay between mass and distance. In considering how gravity changes upon traveling down a thin tunnel to the center of Earth, some students who know the proper relationship between gravity and distance (i.e. they wrote Newton's law of gravitation) reason that gravity *increases* because "the radius is smaller." Perhaps these students assumed that, since Earth's mass is constant, they could consider only distance. They ignore how the relative distribution of mass affects gravity at a point inside Earth. While a truly proper explanation for the decrease of gravity requires a Gaussian approach, students who answered correctly (3.5%, N=5) reasoned that gravity would decrease because, "Some of the mass of the earth would be behind you and pulling in the opposite direction of the rest of the earth," and there is "less mass between you and the center of the body of mass."

Model Mixing

The misconceptions presented above are not necessarily mutually exclusive. For example, a student who asserts that Earth's magnetic field and/or rotation can push out and extend Earth's gravitational field may simultaneously hold the "Boundary Model" misconception and the "Orbital Indicator" misconception. Even students who correctly apply the scientific model may also simultaneously reason with alternative models. It is speculated that students that mix seemingly contradictory models are operating from what Bao and Redish (2006) call a "mixed model state," within which individual models are triggered by situational cues. While seemingly contradictory to the expert, students might use one model in a certain situation and a different model in a different situation. For the students in this study, it was found that, in interview, only when pushed with questions do students begin to question their framework and modify it accordingly.

Despite the multitude of misconceptions about gravity that students can possess, the majority of responses to most of the questions fell within a category that fit with the correct scientific model of Newtonian gravity. However, unlike Piburn (1988), it is not concluded that this means that students mostly have a reasonable understanding of gravity. The misapplications of the scientific model discussed above elucidate the many ways that students get lost in their attempts to reconcile Newton's law of gravitation with their intuition. They learn that mass and distance are the primary factors that determine the strength of gravity, but in many cases they do not know how to apply these factors: Is the "distance" the radius of a planet or the distance of an orbiting body from the surface of a planet? Are mass and distance the *only* factors that determine gravity? How does mass relate to density? It was rare for a single student to possess the correct scientific model without simultaneously relying on an alternative model to answer certain questions. This is in agreement with the results found by Asghar and Libarkin (2010). Perhaps students are trying to "assimilate," rather than "accommodate" in Piagetian terms (Vosniadou and Brewer 1987), what they are taught about gravity with their previous knowledge.

As indicated in the literature, when intuition takes precedence, students often deduce the strength of gravity based on its *effects*, or what one *would feel like in that* *situation.* They have a hard time distinguishing between apparent weight and gravitational force. The "Orbital Indicator" model may be part of a broader mental framework that gravity can be understood in terms of its effect on other objects. For example, student responses that fell in Category 'b' of Question B14 indicate that a small satellite orbiting Earth would experience a *stronger* gravitational force than a larger satellite, because "it is easier to move." Furthermore, responses in Category 'd' of Question A4 (5.6%, N=8) indicate that some students believe a planet's gravity to be defined *only* in terms of its effects on other objects. Responses in this category describe how heavy or light something must be to create its own gravitational field with such responses as: "Heavier than all other objects near it," "Depends upon how big other objects are," and "It would depend on where it is placed. If the object is on earth it would have no gravitational field, but if it were in space it would have one." Determination of gravity based on its effects has been documented by Noce et al (1988) in children and by Asghar and Libarkin (2010) in college geology students. Perhaps this misconception relates to misconceptions about forces in general. Students think in terms of the *effort* it would take to move an object, with the bigger, "stronger" objects being "in control" over the smaller objects.

Discussion

The concept mapping, phenomenographic analysis of student-supplied responses to open-response questionnaires, and characterization of students' mental models described in this chapter represent the Planning phase (Benson & Clark 1982) foundation for constructing a multiple-choice Concept Inventory that can measure Astro 101 students' understanding of Newtonian gravity. Building on the literature of Chapter 2, the most common Astro 101 student responses provide the natural language of students for distractor choices that correspond to the alternative mental models and misapplications of the scientific model discussed above.

CHAPTER FOUR

DEVELOPING THE NEWTONIAN GRAVITY CONCEPT INVENTORY

This chapter discusses the development of a new, valid, reliable multiple-choice assessment instrument to measure Astro 101 students' understanding of Newtonian gravity – the Newtonian Gravity Concept Inventory (NGCI). Informed by the Planning phase of Chapter 3, the three subsequent phases of Benson and Clark's (1982) instrument development model – Construction, Quantitative Evaluation, and Validation – are discussed to provide a full disclosure of the research methods and results used to develop and refine the NGCI iteratively through three versions.

Construction

In constructing the multiple-choice items for the NGCI, the best practices outlined in Haladyna, Downing, and Rodriguez (2002) were implemented. 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 (i.e. those in Table 2 near the beginning of the development process and those in Table 3 near the end of the development process as results helped to inform what the final conceptual domain should be), using typical reasoning errors of students for distractor choices (i.e. the most-common misconceptions uncovered from the open-response questions in Chapter 3), 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.

Pilot Testing

Table 5 shows the pilot testing sites, number of students, and descriptive statistics for all three versions of the NGCI. The first version of the Concept Inventory, simply called the "Gravity Concept Inventory" (APPENDIX I), contained 31 questions, 20 of which were adapted from the open-response questionnaires discussed in Chapter 3. This first version included more questions than the typical Concept Inventory, so that several questions could be piloted and some questions could be eliminated after the statistical analysis. Version 1 was piloted post-instruction during the Spring 2011 Semester with 230 students at Montana State University (MSU). Many students did not finish the survey and had to be eliminated from the sample, and the average score of 38.7% (*SD* = 13.87%) was low, indicating that Version 1 was too difficult and unable to discriminate between students that may have developed a complete and coherent model of gravity from those students that still held inaccurate and incomplete models. Additional review by Physics, Astronomy, and Education experts, analysis of student response patterns (APPENDIX J), and the Quantitative Evaluation discussed in the following section indicated that, as expected, several questions needed to be eliminated or revised to make Version 2 shorter, easier, more readable, more scientifically accurate, and aligned with a clearer conceptual domain. The recommended changes for each item from Version 1 to Version 2 are shown in APPENDIX K.

The Gravity Concept Inventory Version 2 (APPENDIX L) contained 27 questions and was piloted via post-test with 259 MSU students during the Fall 2011 Semester. It showed marked improvement from Version 1. First, the higher average score of 43.4% (*SD* = 18.35%) showed that it was of a more cognitively and developmentally appropriate level. Student response patterns (APPENDIX M) indicated that distractors functioned more effectively, and interviews with seventeen students (students F11-01 through F11-17 in the audio recordings available through APPENDIX C and summarized in APPENDIX N) revealed important information about how students reason when working through the survey. See the Validation section in this Chapter for more information about how these interviews were conducted. During student interviews, it was found that Question 17 contained a typographical error with two choice 'b' options, so student scores for Version 2 were calculated for 26 questions only. Analysis of student interviews, and another round of expert review and Quantitative Evaluation (discussed in the following section) led to specific refinements (see APPENDIX O) for the creation of the third version of the Concept Inventory, the "Newtonian Gravity Concept Inventory" (NGCI).

The NGCI Version 3 (APPENDIX P) contains 26 questions and was piloted during the Spring 2012 Semester both pre- and post-instruction in Astro 101 classes at six institutions from across the United States (Table 5). This includes four large state universities and two community colleges (CC's), each in a different state. 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. In total, 925 Astro 101 students participated in the NGCI pre-test and 743 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). Expert review, interviews with seven Astro 101 students at MSU (students S12-01 through S12-07 in audio-recordings available in APPENDIX C and summarized in APPENDIX R), and Quantitative Analysis (in the section below) indicate that items function as expected and the instrument as a whole is reliable and valid (see the Validation and Discussion sections of this chapter for accompanying interpretive arguments). Therefore, Version 3 represents the final version of the Newtonian Gravity Concept Inventory (NGCI).

Table 5: Astro 101 pilot site data for the NGCI.

Institution	Version	N	Mean %	SD	$\langle g \rangle$	Cohen's *d* (95% C.I.)
Montana State University	1 Post	230	38.71	13.87	n/a	n/a
	2 Post	259	43.42	18.35	n/a	n/a
	3 Pre	278	44.12	17.51	0.09	0.29
	3 Post	245	49.42	19.29		(0.12–0.46)
University of Arizona	3 Pre	287	40.59	17.26	0.24	0.77
	3 Post	237	54.79	19.83		(0.59–0.95)
Westchester CC	3 Pre	72	37.39	16.57	0.55	0.91
	3 Post	66	54.78	21.39		(0.56–1.26)
Youngstown State University	3 Pre	107	36.88	16.36	0.27	0.93
	3 Post	59	53.72	21.05		(0.59–1.26)
Truckee Meadows CC	3 Pre	17	42.08	12.65	0.49	2.12
	3 Post	8	70.67	15.24		(1.03–3.06)
UC Davis Solar System	3 Pre	47	59.00	20.24	0.06	0.13
	3 Post	33	61.54	19.37		(-0.32–0.57)
UC Davis Stars and Galaxies	3 Pre	117	54.65	22.67	0.35	0.75
	3 Post	95	70.66	19.46		(0.47–1.03)

Table 5 also includes the class normalized gains and effect sizes with confidence intervals for Version 3 to provide a rough idea of the impact of the Astro 101 curriculum in each course. One will note that normalized gains, $\langle g \rangle$, are calculated using Hake's (Hake 1998) formula that compares how much students' scores improved to how much they could have improved:

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100\% - \langle S_i \rangle}$$

where $\langle S_i \rangle$ and $\langle S_f \rangle$ represent the class averaged percentage scores pre-instruction and post-instruction, respectively. Cohen's *d* effect sizes with accompanying confidence intervals serve to standardize changes in scores by dividing by the pre- to post-instruction pooled standard deviation, σ_{pooled}, thereby enhancing interpretations of results across institutions (Rodriguez et al 2012 and references therein):

$$d = \frac{\langle S_f \rangle - \langle S_i \rangle}{\sigma_{pooled}}.$$

Gains ranged from 0.06 to 0.55, and effect sizes ranged from 0.12 to 2.12, indicating the NGCI is sensitive to differences in implementation of the Astro 101 curriculum.

Table 6 shows the accompanying Version 3 demographic data calculated as pre-post 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 as per Institutional Review Board approval. Chapter 6 uses Item Response Theory to provide more meaningful measurements of gain and explores how gains are influenced by these student demographics and other classroom dynamics.

Table 6: Astro 101 demographic data for the final version of the NGCI, Verson 3.

		%
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

Quantitative Evaluation

To evaluate the reliability and item functioning throughout each version of the Newtonian Gravity Concept Inventory (Version 1 in APPENDIX I, Version 2 in APPENDIX L, and Version 3 in APPENDIX P), Classical Test Theory (CTT) was used for the Quantitative Evaluation phase of Benson and Clark's (1982) model for instrument development. While CTT is highly sample dependent (Hambleton & Jones 1993, Wallace & Bailey 2010), it is one of the most straightforward and often used statistical methods for evaluating multiple-choice instruments (Ding & Beichner 2009). In particular, it has been used to evaluate both the Star Properties Concept Inventory (Bailey 2006, Wallace & Bailey 2010) and the Light and Spectroscopy Concept Inventory (Schlingman et al 2012), and it provided important information for item revisions to the NGCI throughout the development process (recall that APPENDIX K and APPENDIX O show revisions for each question).

Classical Test Theory

To use Classical Test Theory, items were scored dichotomously, with a correct answer scored as 1, and an incorrect answer scored as 0, so that the number of items answered correctly represents a person, p's, observed score, X_p. Then, CTT posits that this observed score is a reflection of the person's true score, T_p, which is unobservable, but represents what the examinee would have gotten over infinite administrations of the survey, modified by some error, E_p (Lord & Novick 1968, Crocker & Algina 1986, Wallace & Bailey 2010).

$$X_p = T_p + E_p$$

The reliability of the instrument, then, is a measure of how well the variance in observed scores can be explained by the variance of true scores. However, because true scores cannot be measured, the Cronbach's α internal consistency statistic (Cronbach 1951) is often used to place a lower bound on the reliability of an instrument. For K items, α is given by:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right).$$

Values range between 0 and 1 and are highest when the variance of each item ($\sigma_{Y_i}^2$) is small compared to the variance of total test scores (σ_X^2). Note that for dichotomously scored items, the variance of the item is simply $P_i(1 - P_i)$, where P_i is just the proportion of correct responses to the item. Cronbach's α values over 0.70 are conventionally accepted as indicating a reliable instrument (Urdan 2010).

CTT can also be used to comment on the functioning on individual items. 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, item difficulty is reported as the proportion of students that got the item *incorrect*, *D*, so that higher values indicate more difficult items (Schlingman et al 2012):

$$D = \frac{n_0}{n}$$

where n_0 is the number of students who got the item incorrect and n is the total number of students. Questions that everyone gets either correct or incorrect are useless, so 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 (Wallace & Bailey 2010). Ideal *D* values should be halfway between zero and chance for each item (Aubrecht & Aubrecht 1983, Ding & Beichner 2009).

The point-biserial index, r_{pb}, measures an item's ability to discriminate high ability students from low ability students. It is the correlation between examinees' scores on the item and their total scores (Lord & Novick 1968, Crocker & Algina 1986):

$$r_{pb} = \frac{X_1 - X_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

where, again, n_0 is the number of students who answered the item incorrectly, n_1 is the number of students who answered the item correctly, and n is the total number of students. Additionally, X_0 and X_1 are the mean scores of the groups of students who answered the item incorrectly and correctly, respectively, and s_n is the standard deviation of the total group's scores. A value over 0.30 indicates that students' scores on that item are well correlated with their total test scores (Ding & Beichner 2009, Wallace & Bailey 2010).

CTT Results

Table 7 shows Cronbach's α, item difficulty, *D*, and item discrimination, r_{pb}, values for each version of the Concept Inventory. Difficulty values below 0.20 and above 0.80, as well as discrimination values below 0.30, are bolded and italicized. Note that each row represents a single question, and the ordering of items is organized according to those in the final version, Version 4. The corresponding open-response questions are also shown. One can see that Version 1 of the NGCI (APPENDIX I) contained five items (Items 5, 21, 25, 26, and 30) with difficulty values greater than 0.80 while none of the items had values below 0.20, supporting the claim from the Pilot Testing section that the instrument was too difficult. Additionally, ten items (Items 1, 2, 9, 10, 21, 24, 25, 26, 28, and 30, representing 31.3% of the survey) had point-biserials lower than 0.30, indicating that many items were confusing or not aligned with the rest of the instrument (also indicated in the expert review). Four items (21, 25, 26, and 30) that were flagged with a low discrimination index were also flagged as too difficult. Moreover, six of the problem items were found in the last eight items of the instrument, again corroborating the assertion that Version 1 might have been too long and that students were experiencing test fatigue. Therefore, despite an acceptable reliability index of 0.76, the items in Version 1 needed significant changes (see discussions about specific examples in the subsections below and APPENDIX K for all Version 1 item revisions). Eleven items were eliminated for Version 2, while only seven new items were introduced.

Version 2 of the Concept Inventory (APPENDIX L) contained only four items with difficulty values over 0.80 (Item 2, 4, 14, and 18, representing 15.4% of the instrument), and again no Version 2 items were flagged as too easy. However, with only five items (Items 2, 7, 14, 20, and 18, representing 19.2% of the survey) flagged with a low discrimination index, Version 2 showed substantial improvement from Version 1. The higher Version 2 Cronbach's α of 0.79 is additional evidence of improvement. However, Version 2 was still quite difficult for the test population overall. Four items were eliminated for Version 3, three new items were added, and many others were revised (see discussions about specific examples in the subsections below and APPENDIX O for all Version 2 item revisions).

The CTT statistics for Version 3 indicate that these modifications led to further improvement. Two Version 3 items (Items 10 and 25) were too difficult pre-instruction, while only one of these (Item 25) remained too difficult post-instruction. Four Version 3 items (Items 7, 10, 16, and 25) had low discriminatory power for the pre-test, and only two of these (Items 16 and 25) remained poor for the post-test. Additionally, the post-test reliability increased to 0.84, far surpassing the conventionally accepted threshold and precedence of other Astro 101 concept inventories. For comparison, post-test Cronbach's α is 0.75 for the Lunar Phases Concept Inventory (Lindell & 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).

Table 7: CTT statistics: Cronbach's α, item difficulty, and item discrimination for each pilot occasion (pre and post) for each Version (1-3) of the Concept Inventory.

Open Response	Version 1 $\alpha_{post} = 0.76$			Version 2 $\alpha_{post} = 0.79$			Version 3 $\alpha_{pre} = 0.79$, $\alpha_{post} = 0.84$				
	Item #	Post D	Post r_{pb}	Item #	Post D	Post r_{pb}	Item #	Pre D	Pre r_{pb}	Post D	Post r_{pb}
B2	1	0.59	***0.29***	1	0.62	0.36	1	0.69	0.37	0.52	0.47
C10	4	0.77	0.45	3	0.67	0.50	2	0.73	0.44	0.48	0.46
B8	6	0.49	0.44	6	0.45	0.41	3	0.51	0.33	0.37	0.35
	3	0.39	0.37	2	***0.87***	***0.28***	4	0.77	0.49	0.60	0.54
							5	0.30	0.52	0.28	0.50
A1	5	***0.84***	0.42	4	***0.83***	0.36	6	0.49	0.34	0.33	0.37
	10	0.75	***0.22***	7	0.65	***0.07***	7	0.57	***0.21***	0.46	0.34
	11	0.33	0.47	9	0.51	0.60	8	0.71	0.64	0.50	0.61
A6	15	0.38	0.40	8	0.26	0.32	9	0.31	0.31	0.25	0.38
							10	***0.84***	***0.25***	0.79	0.42
	12	0.33	0.45	10	0.37	0.38	11	0.41	0.40	0.27	0.43
	22	0.40	0.46	16	0.46	0.41	12	0.44	0.46	0.35	0.48
B5	18	0.74	0.30	13	0.51	0.50	13	0.64	0.37	0.57	0.46
A8	14	0.68	0.45	12	0.64	0.49	14	0.54	0.48	0.42	0.48
A9	17	0.59	0.53	24	0.56	0.42	15	0.62	0.47	0.46	0.53
				14	***0.83***	***0.23***	16	0.69	***0.20***	0.68	***0.29***
C5	19	0.62	0.43	15	0.51	0.43	17	0.45	0.44	0.37	0.46
B9	24	0.64	***0.23***	20	0.47	***0.24***	18	0.54	0.33	0.39	0.36
				27	0.30	0.50	19	0.41	0.47	0.28	0.47
	20	0.46	0.49	17	n/a	n/a	20	0.46	0.54	0.39	0.53
A4	27	0.29	0.31	26	0.63	0.46	21	0.54	0.39	0.35	0.42
				21	0.43	0.52	22	0.39	0.50	0.32	0.56
C1	23	0.80	0.51	19	0.79	0.47	23	0.52	0.44	0.48	0.44
				23	0.44	0.30	24	0.56	0.32	0.47	0.33
C7	28	0.65	***-0.01***	22	0.74	***0.23***	25	***0.89***	***0.12***	***0.87***	***0.17***
							26	0.40	0.41	0.33	0.49
				5	0.20	0.40					
A3	13	0.55	0.33	11	0.55	0.44					
				18	***0.90***	0.39					
				25	0.40	0.55					
	2	0.45	***0.20***								
	7	0.30	0.39								
B1	8	0.80	0.46								
B10	9	0.69	***0.18***								
B4	16	0.36	0.50								
	21	***0.81***	***0.11***								
	25	***0.85***	***0.25***								
A2	26	***0.86***	***0.02***								
C4	29	0.70	0.46								
B3	30	***0.82***	***0.05***								
	31	0.61	0.49								

Evolution of Questions

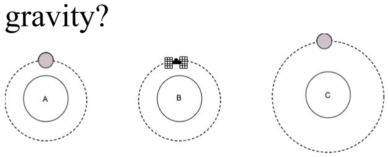
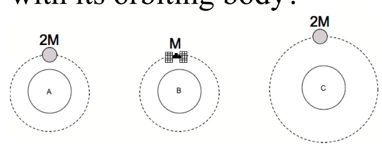
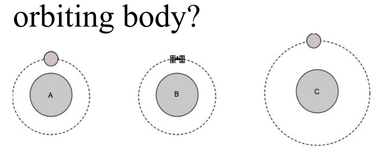
The quantitative analysis done throughout the development of the NGCI indicated problem items for revision and showed that the changes made from version to version led to clear improvement. Following are examples of the specific changes made to the NGCI conceptual focus (resulting in Table 3) and individual items throughout the development process. These are concrete examples of how the best practices for instrument development were implemented, including the results of the CTT quantitative analysis presented in Table 7 as well as suggestions from expert review and student interviews. This discussion is organized according to the conceptual domains outlined in Table 3. Not only does this elucidate how these domains were refined throughout the development process, it grounds the assertion that the intended focus of the NGCI manifests in specific items.

Ontological Categories

Before investigating the evolution of the four concept domains of the NGCI, it is important to first 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 2010), Chi, Slotta, and de Leeuw (1994) and 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 8 below shows an example of how a question on the Concept Inventory Version 1 (originally adapted from open-response question B8) 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 "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, the mass labels were eliminated, and the distractor choices were modified based on Fall 2011 student interviews (students F11-01 through F11-17 in audio recordings in APPENDIX C and summary notes in APPENDIX N).

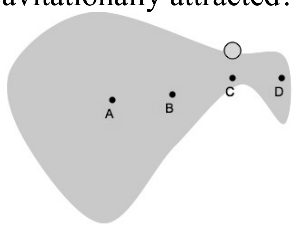
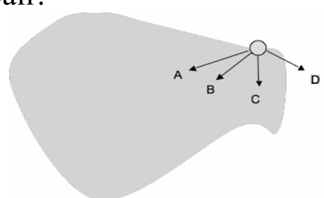
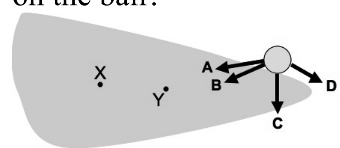
Table 8: An example of an ontological shift from Version 1 to Version 2 of the NGCI. Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is underlined. Difficulty and discrimination values are post-instruction.

Version 1 (Item 6) *D=0.49, r_{pb}=0.44*	**Version 2 (Item 6)** *D=0.45, r_{pb}=0.41*	**Version 3 (Item 3)** *D=0.37, r_{pb}=0.35*
Planets A, B and C have the same mass, but A and C each has a heavy moon orbiting it, and B has a lightweight 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%) <u>All have the same gravity, because they are all the same mass.</u>	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%) <u>Planet A</u> 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.	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%) <u>Planet A</u> b. (8.2%) Planet B c. (8.7%) Planet C d. (13.2%) Both Planets A and B e. (7.1%) All the same

Evolution of the Directionality Domain

Consistent with the initial conceptual focus of Table 2 student ability to reason about the direction of gravity in Versions 1 and 2 of the NGCI was mainly explored with questions about non-spherical objects. Open-ended questions A2 (APPENDIX E) and C9 (APPENDIX G) illustrate the value of this initial qualitative analysis (see the analysis of Chapter 3 and Williamson & Willoughby 2010). An example of a multiple-choice question that remains on the final version of the NGCI is Item 12, the evolution of which is shown below in Table 9. Item 12 asks which direction a ball would roll on an irregularly shaped asteroid. Version 1 used specific points for answer choices, and Version 2 used arrows instead to more directly probe the idea of direction. However, Versions 1 and 2 each required students to know the jargon of “center of mass” and “geometric center.” To reduce this confusion, Version 3 used both the vocabulary and specific points on the diagram. In reviewing the NGCI, however, a few physics experts commented that the correct choice, toward the center of mass, may be correct only to first order, since a completely correct solution requires multipole expansion of the gravitational field. While it is important to maintain scientific accuracy as much as possible in the Concept Inventory, it was decided to leave this question on the final version because it is so useful in discerning students’ conceptions related to the direction of the gravitational force. All three versions of the question have acceptable difficulty and discrimination values, as well as relatively strong distractors, and student interviews suggest this item is functioning as intended (for example, see Interview F11-11).

Table 9: Evolution of Item 12. Percentages of students choosing each choice post-instruction are shown in parentheses, and the correct answer is marked with an asterisk. Difficulty and discrimination values are post-instruction.

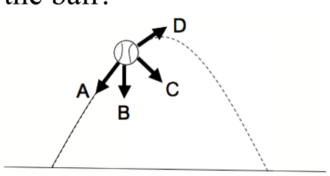
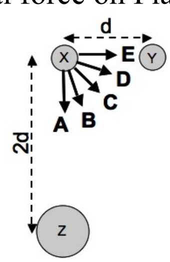
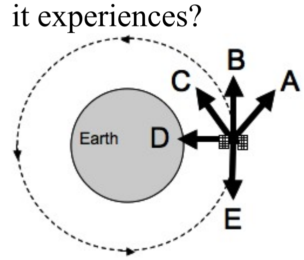
Version 1 (Item 22) *D=0.40,* r_{pb}*=0.46*	**Version 2 (Item 16)** *D=0.46,* r_{pb}*=0.41*	**Version 3 (Item 12)** *D=0.35,* r_{pb}*=0.48*
Consider a small, lumpy asteroid. Toward which point will the ball be gravitationally attracted?  a. (59.6%) Point A, because it is the center of mass of the asteroid. b. (7.8%) Point B, because it is the geometric center of the asteroid. c. (28.7%) Point C, because it is directly below the ball. d. (3.9 %) Point D, because it is at the edge nearest the ball.	Consider an irregular-shaped asteroid of constant density shown below. Which arrow best represents the direction of the gravitational force on the ball?  a. (54.7%) A, because it points to the center of mass of the asteroid. b. (25.1%) B, because it points to the geometric center of the asteroid. c. (13.1%) C, because it points straight down to directly below the ball. d. (7.1%) Point D, because it is at the edge nearest the ball.	For the asteroid shown below, X represents the center of mass and Y represents the geometric center. Which arrow best represents the direction of the gravitational force on the ball?  a. (65.4%) A, because it points to X. b. (16.4%) B, because it points to Y. c. (15.4%) C, because it points beneath the ball. d. (2.9%) D, because it points down the slope.

While student responses to questions about the direction of gravity for non-spherical objects are informative, they can be somewhat limited in terms of exploring the full range of students' ideas related to the direction of the gravitational force. Therefore, for Version 3 of the NGCI an effort was made to include more questions that explore student understanding of the direction of force with relative motion and superposition. In fact, all three of the questions introduced in Version 3 were designed to probe the Directionality domain. Table 10 shows these three additional directionality questions.

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 (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 associated with 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. This is likely 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." This idea also manifests in student responses to Item 6 discussed within the Evolution of the Threshold domain section below.

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, supporting the results of Dostal (2005). Interviews also suggest that the remaining distractors are chosen by students who reason with intuition rather than proportional reasoning (for example, see Interviews F12-03 and S12-02). 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. Table 7 shows that, despite its difficulty, this question is actually a very good discriminator of student ability.

Table 10: 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.

Version 3 (Item 5) *D=0.28, r_{pb}=0.50*	**Version 3 (Item 10)** *D=0.79, r_{pb}=0.42*	**Version 3 (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?  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?  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?  a. (2.0%) A b. (6.1%) B c. (19.7%) C d. <u>(67.3%) D</u> e. (3.4%) E

Evolution of the Force Law Domain

Originally, the open-response questions and the questions in Version 1 of the NGCI related to the Force Law domain were not intended to directly address proportionality within the force equation beyond simple increase/decrease relationships (i.e. if the distance increases the force decreases). However, several of the interactive engagement instructional strategies for use in Astro 101 classrooms, such as *Lecture-Tutorials for Introductory Astronomy* (Prather et al 2012) and Ranking Tasks (Hudgins et al 2006), 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). To assess how frequently Astro 101 instructors require their students to do this, the AstroLrnr listserv and 60 Physics and Astronomy Departments from around the United States were emailed. 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, it was deemed 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. Below are two examples of this change to items in the Force Law domain.

Table 11 shows how Version 1 Item 3 was worded so that a student would simply need to remember that the gravitational force diminishes *quickly* or *drastically* to get the correct answer, rather than implement mathematical proportional reasoning. The choices in Version 2 of the question (Version 2 Item 2), however, are much more explicit about exactly *how much* the gravitational force changes. The large factor of ten from Version 1 was no longer needed to probe this concept, so the factor was simply changed to two.

However, one will notice that the correct answer is ‘d. None of the above is correct.’ Interviews (such as F11-09) indicated that students did not consider ‘None of the above’ when presented with a distractor as effective as choice ‘a,’ making the question overly difficult and a poor discriminator of student ability as measured by the CTT statistics. Additionally, one can see from Table 11 that choices ‘b’ and ‘c’ were not strong distractors. To address these issues, the distance factor for Version 3 (Item 4) was chosen to be four so that each choice option could be a clear numerical progression of ‘weaker.’ As seen in Table 7, these changes were effective in creating an item of appropriate difficulty and discriminatory power. Again, as in Dostal (2005), student interviews overwhelmingly indicate that the most effective distractor is that which represents the misconception that distance, rather than distance squared, is inversely related to gravitational force.

The most difficult item on the final version of the NGCI, Item 25, also falls within the Force Law domain, and as seen in Table 7, this question maintained a high CTT difficulty value throughout the development process, and Table 12 shows its evolution through drastic changes. The essential purpose of each multiple-choice version of this item is to probe student understanding of how gravitational force is affected by a planet’s mass and size. Versions 1 and 2 use Earth and Saturn as a real-life comparison of planets that are very different in size yet similar in the gravitational force they exert on objects on their surfaces. Note the shift from “surface gravity” in Version 1 to “your weight” for Version 2 as a way of addressing the ontological framing issue. Additionally, the

Table 11: The evolution of Item 4 on the 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 **Item 3** *D=0.39,* r_{pb}*=0.37*	**Version 2** **Item 2** *D=0.87,* r_{pb}*=0.28*	**Version 3** **Item 4** *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? a. (7.4%) Would be unchanged b. (17.0%) Would decrease only slightly c. (14.8%) Would be half as strong d. <u>(60.9%) Would decrease a huge amount.</u>	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? 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. <u>(13.1%) None of the above is correct.</u>	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. (7.1%) Two times weaker. b. (35.7%) Four times weaker. c. (16.8%) Eight times weaker. d. <u>(40.1%) Sixteen times weaker.</u>

Version 2 distractor choice 'b' draws on a misconception uncovered in open-response Question C7 (APPENDIX G), which probes whether students account for a planet's rings when reasoning about gravity. However, student interviews revealed a variety of unintended issues with this item. Some students were confused because it would be impossible for a person to stand on the cloudy surface of Saturn, and some students simply did not believe what the question was asserting (see Interview F11-06). To test the same basic idea without relying on students' knowledge of Saturn, the question was converted to a completely hypothetical situation for Version 3. At first encounter, the Version 3 question looks very similar to two other items on the NGCI Version 3 (Items 14 and 17), where simple reasoning about mass, rather than size (or distance), can lead to the correct answer. 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. Here, the most effective distractor corresponds to the idea that "same gravity implies same mass," which is only true if distance is the same. So, even students that perform very well on the NGCI may get this question wrong because they too quickly jump to this idea without double checking the factor of distance, an this likely explains the high difficulty and poor discrimination index of this question.

Table 12: The evolution of Item 25 on the 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 **Item 28** *D=0.65,* r_{pb}*=-0.01*	**Version 2** **Item 22** *D=0.74,* r_{pb}*=0.23*	**Version 3** **Item 25** *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 …  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.

However, it is vitally important for students to be able to correctly distinguish situations where either mass or distance is different. Item 25 is well aligned with the concept domains of the NGCI presented in Table 3, interviews (such as S12-02) suggest it functions as intended, and removing it from the analysis does not increase the reliability of the NGCI, so, despite its poor performance in terms of CTT statistics, it is concluded that Item 25 should remain on the final NGCI.

Evolution of the Independence of Other Forces Domain

As discussed in the literature review of Chapter 2, one of the most prevalent and well-documented student misconceptions concerning the force of gravity is that it is confounded by other forces such as magnetism, rotational forces, or forces due to air pressure (recall, for example, Nussbaum & Novak 1976, Treagust & Smith 1989, Bar & Zinn 1998, and Asghar & Libarkin 2010). To probe this robust alternative model, the Independence of Forces domain focuses on how the gravitational force relates to these other forces. Because the multiple-choice questions that were created for Version 1 of the NGCI were constructed with robust information from student-supplied descriptions of the effects of other forces characterized in the Planning phase discussed in Chapter 3 and in Williamson and Willoughby (2012), they generally functioned as expected. Therefore, changes to these questions throughout the development of the NGCI tended to be minor.

Table 13 shows one example of a question (Item 8) which was changed to improve its performance and better align its answer choices with the Independence of Other Forces Concept domain. Note the ontological shift in the phrasing of the stem from Version 1 to Version 2. Furthermore, by removing answer choices that involve density and size, the question more accurately probes students' 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.

Table 13: The evolution of Item 8 on the 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, Item 11 *D=0.33, r_{pb}=0.47*	**Version 2, Item 9** *D=0.51, r_{pb}=0.60*	**Version 3, Item 8** *D=0.50, r_{pb}=0.61*
Why does the Earth have gravity? a. <u>(67.4%) It has mass.</u> b. (10.9%) It has a magnetic field. c. (5.7%) It is so large. d. (6.5%) It is so dense. e. (9.6%) It rotates. f. (0.0%) It has an atmosphere.	Why does the Earth exert a gravitational force on objects on its surface? a. (17.6%) It has an atmosphere b. (3.7%) It is very dense. c. (16.5%) It has a magnetic field. d. <u>(48.3%) It has mass.</u> e. (13.9%) It rotates.	Why does Earth exert a gravitational force on objects on its surface? a. (3.4%) It has an atmosphere. b. (5.3%) It has a magnetic field. c. <u>(49.5%) It has mass.</u> d. (2.2%) It rotates. e. (39.4%) More than one of these.

<u>Evolution of the Threshold Domain</u>

The items of the NGCI that are aligned with the Threshold Concept domain probe a combination of students' alternative mental models that were identified in Chapter 3 related to the universality of gravity in 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). Table 14 and Table 15 show the evolution of two Version 3 items (Items 21 and 6) to illustrate how students' ideas about these thresholds can be inextricably linked to observed motion.

Item 21 (Table 14), requires students to reason about the force between Earth and a list of specific objects. While the original open-response question (Question A4 in APPENDIX E) and the Version 1 question 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, such as in Interview F11-10). 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, & 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.

Another idea in the Threshold domain is the common misconception that an atmospheric threshold must exist as a condition for orbiting and weightlessness. Recall that 50.3% of the students who responded to the open-response question A1 (APPENDIX E) indicated that astronauts appear to float in their spacecraft because there is no gravity in space. The results from previous studies (Ruggiero et al 1985, Sharma et al 2004) and student interviews demonstrate that this misconception often is tied closely to the presence of an atmospheric boundary. To many students, ‘space’ generally means outside of Earth’s atmosphere. However, even for students who do understand that objects in space can gravitationally attract, this question is particularly difficult because they do not readily have an explanation for the contradiction of astronauts in orbit who are apparently weightless.

Table 14: The evolution of Item 21 on the 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.

Open-Response **A4**	**Version 1** **Item 18** *D=0.29,* r_{pb}*=0.31*	**Version 2** **Item 13** *D=0.63,* r_{pb}*=0.46*	**Version 3** **Item 13** *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. (71.7%) Any mass will do 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. (36.7%) All of them	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. (65.1%) All of them

The evolution of NGCI Version 3 Item 6 (Table 15), which asks about why an astronaut floats in a spacecraft, reflects this issue. Version 1 of this question proved to be too difficult for students, because it required students to understand that orbiting is "falling." For many students, however, the word "falling" *only* means falling *down*. In an attempt to eliminate this hurdle in language, Version 2 described the astronauts as "moving at the same speed as their spacecraft" instead. However, this change in language did not reduce confusion or the difficulty of this item. In interviews, students were unable to connect the idea that moving at the same rate could cause weightlessness, as there are plenty examples of people on Earth moving at the same rate as a car or plane who do not experience weightlessness (for example, see Interviewee F11-01). So, Version 3 of the question was simplified to address the gravitational force directly rather than *apparent weightlessness*. Not only is this more straightforward in terms of students' natural language, it aligns better with the recommendations in the literature (Galili 1995 and 2001, Galili & Lehavi 2003) and more directly probes the idea of 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 (Table 3). The CTT statistics shown in Table 3 show that these changes were effective in creating an item of appropriate difficulty and discrimination.

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. This section outlines assumptions with supporting evidence to argue that the NGCI is a valid tool for measuring Astro 101 students' understanding of Newtonian gravity.

Table 15: The evolution of Item 6 on the 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.

Open-Response **A1**	**Version 1** **Item 5** *$D=0.84$, $r_{pb}=0.42$*	**Version 2** **Item 4** *$D=0.83$, $r_{pb}=0.36$*	**Version 3** **Item 6** *$D=0.33$, $r_{pb}=0.37$*
Why do astronauts appear to float in their 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 free-fall. • 2.1% The spacecraft's gravity isn't strong enough.	Astronauts appear weightless in their spacecraft because… 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.	Astronauts appear weightless in their Earth-orbiting spacecraft because: 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.	An astronaut floating in her Earth-orbiting spacecraft… 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.

The first and primary validity assumption is that the NGCI adequately encompasses the tenants of the Newtonian gravity construct and that students' performance on the NGCI indicates their proficiency within this construct. In Chapter 3 and throughout this chapter, a variety of sources were used to establish the most important aspects of the Newtonian gravity construct (see Table 1, Table 2, and Table 3) and to integrate these into the design of questions (for both the open-response questionnaires and Versions 1-3 of the multiple-choice Concept Inventory). Table 3 in

Chapter 3 shows the final four concept domains that clearly define the scope of the final version 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 are shown in Table 16 below. The Fleiss' Kappa inter-rater reliability statistic for this expert categorization 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, the Evolution of Questions section of this chapter shows how ten 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 MSU Physics faculty and graduate students evaluated the correctness of items by providing their answers to NGCI Version 3. Their average score was 96.7%. Additionally, as measured by the CTT statistics with the test population (Table 7), 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.

Table 16: Expert categorization of NGCI items by Doman. Each cell shows the number of raters (out of 4) who marked which item belongs in which concept domain.

Item	Directionality	Force Law	Other Forces	Threshold
1			4	
2	3			1
3		4		
4		4		
5	4			
6				4
7		4		
8			4	
9		2	2	
10	4			
11		4		
12	4			
13		2		2
14		4		
15			4	
16		4		
17		4		
18		4		
19		4		
20		2		2
21				4
22		1	3	
23				4
24				4
25		2	2	
26	4			

Further, the wide range of observed normalized class gains (0.09–0.55) and effect sizes (0.13–2.12) (Table 5) indicate 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 along a representative Newtonian gravity construct.

The second assumption of validity asserts that NGCI questions are reasonable to the test population and, consequently, students interpret the NGCI as intended. Specifically, it is asserted that students understand what each question is asking and that distractor choices function effectively as probes of student misconceptions. There are two primary pieces of evidence to support these claims. First, student-supplied responses to the open-ended questionnaires (APPENDIX E, APPENDIX F, and APPENDIX G) during the Planning phase of development (Chapter 3) 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. One can gauge the effectiveness of these distractors by looking at the distribution of student responses (APPENDIX J, APPENDIX M, and APPENDIX Q). There are additional discussions for the ten specific example items in the Evolution of Questions section.

The second piece of evidence to support the validity assumption includes information from interviews with students who worked through the multiple-choice questions of the Concept Inventory (APPENDIX C, APPENDIX N, and APPENDIX R). 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 (APPENDIX C and APPENDIX N), and seven were interviewed while taking Version 3 (APPENDIX C and APPENDIX R). 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 (APPENDIX C) 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 & Bronkhorst 1996), interrupting the students only to remind them to clarify or elaborate. The researcher took notes during the interview and elaborated on relevant issues as soon as possible after the interview was over (APPENDIX N and APPENDIX R). As discussed in the Evolution of Questions section, 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 (for example, see the discussion of Version 2 Item 23 with Interviewee F11-16). Finally, interviews with students who were taking Version 3 of the NGCI (interviews S12-01 through S12-07) 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 help to establish the NGCI as a valid instrument in measuring Astro 101 student understanding of Newtonian gravity. The NGCI adequately encompasses the tenants of the Newtonian gravity construct, students' performance on the NGCI indicates their proficiency within this construct, the NGCI questions are reasonable to the test population, and students interpret the NGCI as intended.

Discussion

This chapter makes explicit the survey design and development process of the Newtonian Gravity Concept Inventory (NGCI) for measuring Astro 101 student understanding of Newtonian gravity within four conceptual domains: (1) Directionality, (2) Force Law, (3) Independence of Other Forces, and (4) Threshold. The methods for the instrument development process followed the best practices outlined in Benson & Clark (1982). First, the Planning phase involved reviewing previous work, defining a conceptual focus, and performing a phenomenographic and interpretive analysis of student interviews and written answers to open-response questionnaires to characterize Astro 101 student understanding of Newtonian gravity (Chapter 3). Second, the Construction phase involved adapting open-response questions into multiple-choice format by using the most common student-supplied responses as distractor choices, as well as creating new questions and revising iteratively with feedback from student answer choice distributions, student interviews, expert review, and Classical Test Theory. This CTT analysis represented the third phase of Quantitative Analysis, and it showed that the NGCI is reliable with items that are of adequate difficulty (neither too hard nor too easy) and can discriminate between Astro 101 students with varying degrees of understanding of Newtonian gravity. Finally, the Validation phase outlined two assumptions that underlie interpretations of NGCI scores and supported these assumptions with various lines of evidence. It is concluded that the final version of the NGCI, Version 3 (APPENDIX P), is a robust, reliable, and valid instrument for measuring Astro 101 student understanding of Newtonian gravity.

CHAPTER FIVE

USING THE NGCI TO COMPARE PHYSICS AND ASTRONOMY STUDENTS

As discussed throughout Phase I in this thesis, analyzing Astro 101 student performance on the NGCI offers important insight into the difficulties this population experiences when learning and reasoning about Newtonian gravity. Furthermore, it is important to remember that, within the Constructivist framework, these difficulties manifest in relation to the typical Astro 101 student's background experience as well as the typical Astro 101 classroom environment and curriculum. Most Astro 101 students are non-science majors who have little previous experience with math or physics courses. These students take the course as a one-time general education science requirement before going on to become society's businesspeople and teachers (Franknoi 2002, Prather, Rudolph, & Brissenden 2009). Additionally, most Astro 101 courses investigate gravity within the context of tides, planet and star formation, space travel, and to provide context for the structure of the universe. Accordingly, this research is an important extension of the results of other Astro 101 concept inventory investigations into student learning of topics related to lunar phases (Lindell & Olsen 2002), light and spectroscopy (Bardar et al 2007), star properties (Bailey 2006), and the solar system (Hornstein et al 2011). However, unlike these other Astro 101 topics, the topic of Newtonian gravity is also taught in a typical introductory algebra-based physics course, hereafter "Phys 101" or just "physics." The Phys 101 population represents a vastly different group than the Astro 101 population in terms of interest, motivation, comfort with math and science, and career path. Most physics students are science (or engineering, technology, and pre-professional track) majors and have had previous instruction in astronomy or physics. Also, the typical Phys 101 course investigates Newtonian gravity within a very different context than the Astro 101 course, with problems and examples that tend to focus on the effects of gravity on or near Earth's surface, especially in calculating weight forces and mainly in context with Newton's Laws. Indeed, much of the literature on student understanding of gravity has focused on physics students (recall Chapter 2).

It is a natural progression of this study to use the NGCI to investigate the differences between the two very different populations of Astro 101 and Phys 101 students. Therefore, Phase II of this thesis attempts to answer the following questions:

- Does the NGCI's ability to reliably and validly measure Astro 101 student understanding of Newtonian gravity extend to the Phys 101 population?
- Do the Astro 101 and Phys 101 populations approach the NGCI differently? If so, how?
- Do the Astro 101 and Phys 101 populations have different levels of ability along the construct of Newtonian gravity?
- Do the differences in Astro 101 and Phys 101 students' interests and backgrounds and classroom curricula differentially influence students' learning of Newtonian gravity? If so, how?

With the addition of Phys 101 pilot site testing data, this chapter addresses the first two questions with an analysis of Classical Test Theory statistics and comparisons of student response patterns. Chapter 6 addresses the last two questions with Item Response Theory and Multiple Linear Regression models with student demographic and classroom dynamics data.

Applicability of the NGCI to Phys 101

The Newtonian Gravity Concept Inventory was developed with and for the Astro 101 student population, and, while it may seem reasonable that it would extend to the Phys 101 population, it needs to be tested. To answer the question of whether the NGCI can provide reliable and valid measures of Phys 101 student understanding of Newtonian gravity, and it must undergo additional pilot testing, quantitative analysis, and validation.

Phys 101 Pilot Sites

During the Spring Semesters in 2012 and 2013, the NGCI was piloted in Introductory Algebra-Based College Physics classes at eight colleges and universities in the United States. Table 17 provides descriptive statistics, as well as normalized gain and effect size where appropriate, for these pilot sites. In total, 1,392 Phys 101 students participated in the NGCI pre-test, with a population average score of 57.88%, and 929 Phys 101 students participated in the post-test, with a population average score of 67.86%. Average class pre-test scores ranged from 43.42% to 71.21%, and post-test scores ranged from 50.14% to 85.10%. Gains ranged from 0.12 to 0.52. As with the Astro 101 pilot testing data, these broad ranges suggest that the NGCI is indeed sensitive to differences in Phys 101 instruction.

Table 18 shows the accompanying Phys 101 demographic data. As with the Astro 101 sample, these percentages were calculated as pre-post averages of the students who responded to the demographic questions at the end of the NGCI, and students who reported an age of 17 or younger were eliminated from the sample.

Table 17: Introductory Physics pilot site data for the NGCI.

Institution		N	Mean %	SD %	$\langle g \rangle$	Cohen's *d* (95% C.I.)
Montana State University	2012 Post	73	59.80	16.54	n/a	n/a
	2013 Pre	287	51.49	18.9	0.19	0.51
	Post	218	60.78	17.18		(0.33-0.69)
College of Dupage	2012 Pre	49	67.74	18.74	0.52	0.96
	Post	33	84.50	15.28		(0.49-1.42)
Northern Arizona Univ.	2012 Pre	65	49.23	15.46	0.28	0.74
	Post	58	63.39	22.39		(0.37-1.10)
	2013 Pre	184	43.42	17.48	0.12	0.38
	Post	157	50.14	18.18		(0.16-0.59)
University of Maine	2012 Post	123	59.88	19.44	n/a	n/a
	2013 Pre	245	57.74	18.69	0.36	0.86
	Post	131	72.99	15.62		(0.64-1.08)
UC Santa Barbara	2013 Pre	367	71.21	19.41	0.48	0.80
	Post	221	85.10	13.31		(0.63-0.97)
Buffalo State University	2013 Pre	29	53.18	19.57	0.37	0.83
	Post	21	70.33	21.83		(0.24-1.41)
Snow College	2013 Pre	20	53.46	19.13	0.40	0.98
	Post	17	71.94	18.73		(0.27-1.64)
Kilgore College	2013 Pre	23	44.65	16.27	n/a	n/a

Table 18: Phys 101 demographic data for the NGCI.

		%
Major	Business	0.8
	Education	2.0
	Humanities, Social Sciences, Arts	3.2
	Science, Engineering, Architecture	87.4
	Other	6.5
Previous Courses	0	39.7
	1	42.7
	2 or more	17.4
Gender	Male	60.2
	Female	39.5
Age	18-30	97.9
	Older than 30	1.9

CTT Analysis of the NGCI for Phys 101

Because Classical Test Theory statistics are highly sample dependent, they must be recalculated for each group of students. Therefore, the NGCI will have a different reliability index and different item difficulty and discrimination parameters for physics students than it does for astronomy students. The CTT statistics for the Physics sample are shown in Table 19. One can see that the Phys 101 Cronbach's α is very high, actually higher than the Astro 101 values both pre-instruction and post-instruction, indicating that the NGCI is a reliable instrument for the Phys 101 population. Additionally, the items on the NGCI are generally of appropriate difficulty and discriminatory power for Phys 101 students. One item (Item 25) is too difficult pre-instruction and one (Item 5) is too easy. No items are too difficult post-instruction, however, four items (Items 5, 19, 20, and 26) are too easy. Because 97% of physics students answer Item 5 correctly post-instruction, it is also not a good discriminator of student ability. However, all other items are have good discrimination values both pre and post-instruction.

Validation

While the NGCI was not developed with information from Phys 101 students' responses, many of the same arguments from Chapter 4 that support the NGCI as a valid tool for measuring Astro 101 students' understanding of Newtonian gravity apply. First, the construct of Newtonian gravity as elucidated in Chapters 2 and 3 is taught in both Astro 101 and Phys 101 courses. Students in both courses are expected to understand Newtonian gravity within all four of the conceptual domains in Table 3. Indeed, the extant literature review presented in Chapter 2 actually emphasizes the misconceptions of physics students more so than astronomy students. Second, since the development of the NGCI was strongly informed by the literature with insight from both physics and astronomy experts, and since it was actually a group of physics experts who evaluated the correctness of answers, it is more than reasonable to assume that the NGCI is an appropriate instrument for probing Phys 101 students' understanding of gravity.

Table 19: NGCI CTT statistics for Phys 101: Cronbach's alpha, item difficulty, and item discrimination both pre- and post-instruction. Difficulty values below 0.20 and above 0.80, as well as discrimination values below 0.30, are bolded and italicized.

Version 3 $\alpha_{pre} = 0.84$, $\alpha_{post} = 0.86$				
Item	Pre *D*	Pre r_{pb}	Post *D*	Post r_{pb}
1	0.57	0.39	0.48	0.44
2	0.47	0.51	0.29	0.56
3	0.39	0.31	0.32	0.42
4	0.63	0.54	0.48	0.59
5	***0.13***	0.39	***0.03***	***0.21***
6	0.36	0.38	0.26	0.40
7	0.52	0.36	0.46	0.50
8	0.51	0.67	0.38	0.66
9	0.26	0.30	0.21	0.40
10	0.77	0.49	0.65	0.60
11	0.30	0.41	0.22	0.40
12	0.27	0.42	0.24	0.51
13	0.45	0.45	0.35	0.44
14	0.42	0.51	0.28	0.50
15	0.51	0.55	0.35	0.51
16	0.66	0.39	0.55	0.49
17	0.36	0.47	0.26	0.48
18	0.41	0.33	0.28	0.37
19	0.22	0.40	***0.18***	0.39
20	0.25	0.53	***0.19***	0.51
21	0.34	0.49	0.20	0.45
22	0.31	0.57	0.23	0.59
23	0.32	0.50	0.25	0.44
24	0.44	0.41	0.35	0.36
25	***0.82***	0.34	0.71	0.42
26	0.25	0.48	***0.14***	0.42

Third, the following sections of this chapter show that the NGCI is capable of providing valuable, nuanced information about physics students' reasoning patterns that are slightly different from astronomy students'. And finally, it is noted that some of the Astro 101 interviewees took physics before or concurrent with their astronomy course (such as F11-01, F11-02, F11-12, F11-13, F11-17, S12-06, and S12-07), giving insight into how physics students approach the NGCI, with no indication that the NGCI was less appropriate for them than the rest of the Astro 101 students. To be sure, additional interviews conducted with eleven MSU physics students[8] (students F12-01 through F12-11 in the audio recordings in APPENDIX C and notes in APPENDIX S) corroborate that physics students interpret NGCI items and distractor choices as intended.

In light of the evidence above, it is asserted that the NGCI *can* provide a reliable and valid measurement of introductory algebra-based physics students' understanding of gravity. The instrument is sensitive to a range of instruction in physics classes, most NGCI items are of appropriate difficulty and discriminatory power for physics students, although there are some notable problem items, and distractor choices are effective in eliciting known misconceptions (as discussed below).

Comparing Astro 101 and Phys 101 Using CTT and Response Patterns

As mentioned above, the NGCI is sensitive to nuanced differences in both Astro 101 and Phys 101 student understanding of Newtonian gravity. What kind of

[8] Interviews were with second-semester algebra-based physics students just after they had finished Phys 101.

comparisons can be made between these two very different populations? Just from the descriptive statistics in Table 18 and Table 19, one can already see emerging differences. First, the Phys 101 group averaged pre-instruction and post-instruction scores were 14.17% and 12.47% *higher* than that of the Astro 101 group, respectively. Average class pre-instruction scores ranged from 37% to 59% for the astronomy classes, but from 43% to 71% for the physics classes. Post-instruction class scores ranged from 49% to 70% for the astronomy classes, but from 50% to 85% for the physics classes. The demographics of the samples reflect the trends discussed above as well: 67.4% *more* of the Phys 101 sample reported a major of "Science, Engineering, or Architecture," and 17.4% *more* of the Phys 101 sample had taken at least one previous astronomy or physics course. As with the descriptive statistics in Table 17 and Table 18, the CTT statistics already hint at differences between Astro 101 and Phys 101 students. The mean pre and post-instruction item difficulty values calculated with the Phys 101 sample were 0.42 and 0.32, which were 0.14 and 0.13 lower than that for the Astro 101 sample, corroborating the argument that the NGCI is somewhat easier for the physics group. The mean pre- and post-instruction physics item discrimination values were 0.45 and 0.46, which were only 0.06 and 0.02 higher than those for the astronomy sample. So, overall, NGCI items function equally well in discriminating between students of low and high ability for both astronomy and physics students. The significance of these differences will be explored in Chapters 5 and 6.

To get a better idea of how CTT item parameters vary by group (i.e. physics or astronomy), Figures 4 and 5 plot post-instruction values against pre-instruction values.

With this representation, the dotted line represents no change in the parameter, so items that fall below the dotted line in Figure 4 became less difficult after instruction, and items that fall above the dotted line in Figure 5 became more discriminating. One can see that for both astronomy and physics, most items became less difficult and more discriminating, indicating that the NGCI measures student learning in both populations. However, the clear shift in the distributions of physics and astronomy difficulty values in Figure 4 shows that NGCI items are generally easier for physics students than astronomy students. Figure 4 again shows that, despite their difference in difficulty, items on the NGCI have generally the same distribution in discriminatory power for both populations, but eight items decrease in discrimination for physics whereas only three items decrease in discrimination for astronomy. This may be an effect of the background of the physics sample. Many more physics students had previously taken a physics or astronomy course, so perhaps the higher pre-instruction discrimination parameters are simply discriminating between students who have seen the curriculum before and those who have not.

Differences in the overall functioning of the NGCI between astronomy and physics students can be probed by further exploring the CTT parameters for both groups and comparing the student response patterns to gauge the functioning of distractor choices to individual items (APPENDIX Q). Through this investigation, interpretive arguments about the differences between how physics and astronomy students' understanding of Newtonian gravity within each of the four conceptual domains of the NGCI (Table 3) are supported and explored with specific examples in the subsections below.

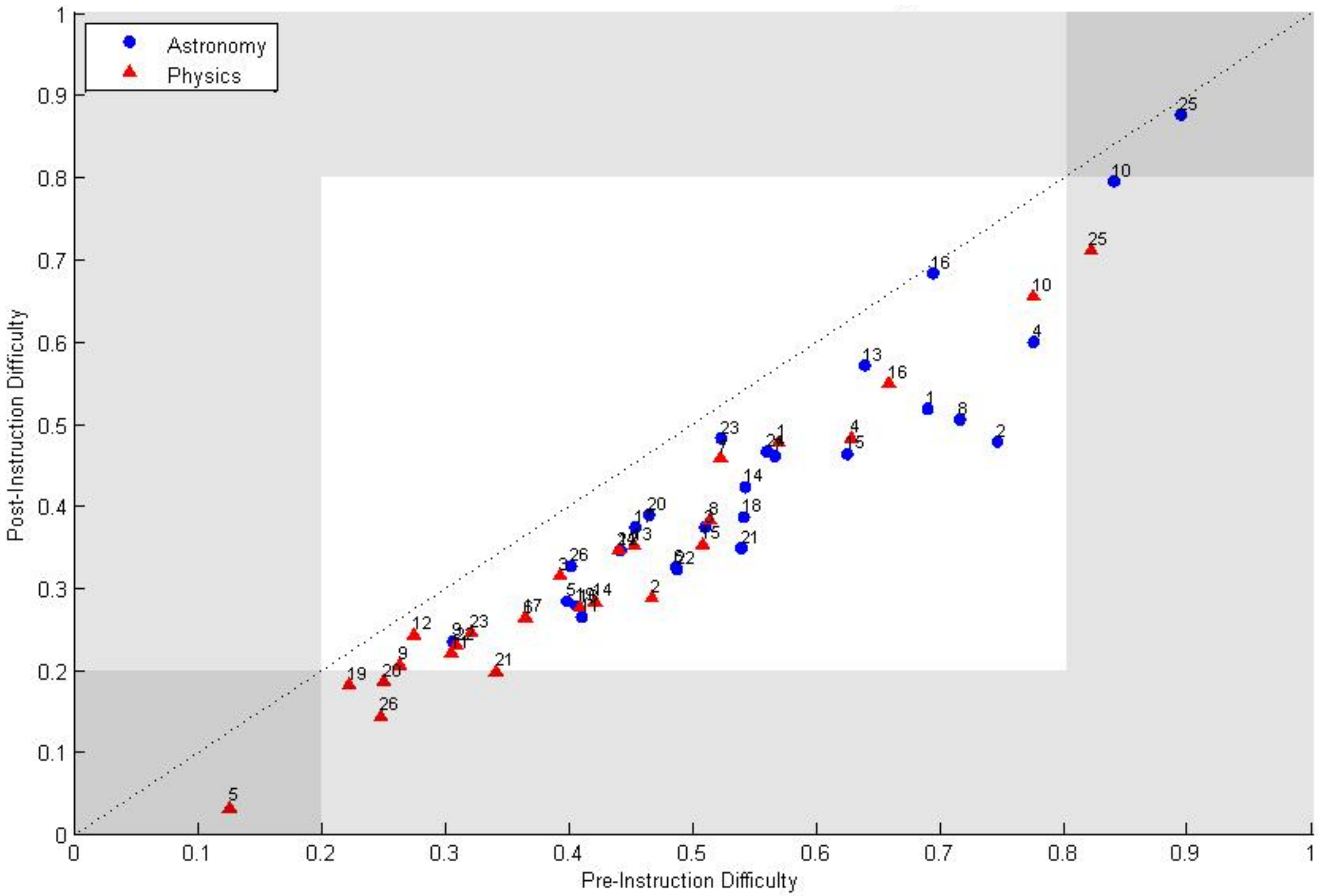


Figure 4: Post-Instruction versus pre-instruction CTT difficulty values calculated with the physics and astronomy samples. Data points are labeled with their corresponding item number.

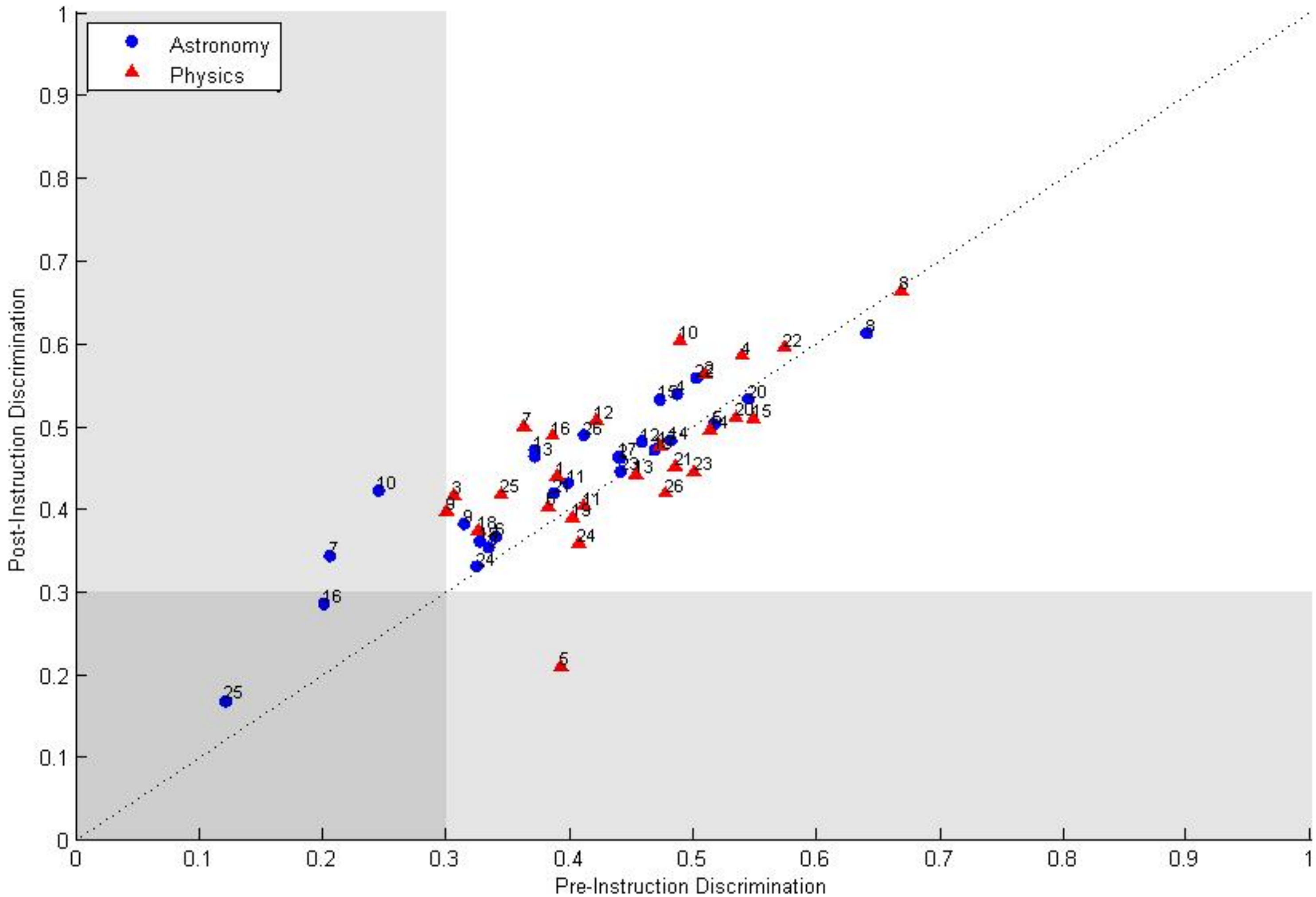


Figure 5: Post-Instruction versus pre-instruction CTT discrimination values calculated with the physics and astronomy samples. Data points are labeled with their corresponding item number.

Differences in the Directionality Domain

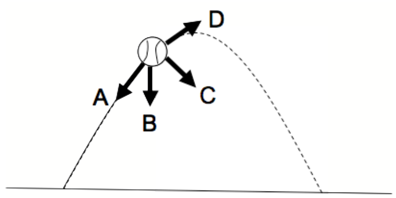
Differences between how physics and astronomy students reason about the direction of the gravitational force are best captured by differences in student response patterns to Items 5, 12, and 26 (see Table 20). As discussed in Chapter 4, Items 5 and 26 are conceptually identical, and physics students perform much better than astronomy students on both items, but Item 5, which asks about the direction of the gravitational force on a ball thrown in the air near Earth's surface, functions most differently (compared to all other items) in physics and astronomy CTT difficulty values (i.e. it has the largest distance between physics and astronomy CTT data points in Figure 4). While it is an appropriate difficulty and good discriminator for the astronomy sample, it is far too easy for physics students both pre- and post-instruction, which likely explains why it is also a poor discriminator for the physics sample post-instruction. Item 26, however, is too easy for physics students only post-instruction (and not nearly as easy as Item 5), and it is consistently a good discriminator of student ability. Therefore, even though these items are conceptually similar, *both* populations are answering Item 5 differently from Item 26. Therefore, *both* populations are distracted by the surface features of the items (Chi, Feltovich, and Glaser 1981). What kinds of misconceptions do the surface features trigger in students' mental landscapes? First, students who hold the Boundary Model misconception may approach Item 26 differently than Item 5 because they would account for the presence of Earth's atmosphere as a boundary, or "check point," to reason that objects outside of Earth's atmosphere (i.e. in space) behave differently than objects inside of Earth's atmosphere with regard to gravity (i.e. the "Earth is special" misconception discussed in the literature review of Chapter 2). Additionally, students who choose option ‘c’ in both questions generally reason with the idea that gravitational force is in the direction of motion. Table 20 shows that this is the most chosen distractor for *both* astronomy and physics students, and the other distractor choices function similarly. So, even though most physics students have a much better overall understanding of the direction of the gravitational force (i.e. they choose distractors less often and therefore have higher scores on the NGCI), some still hold the same misconceptions as astronomy students.

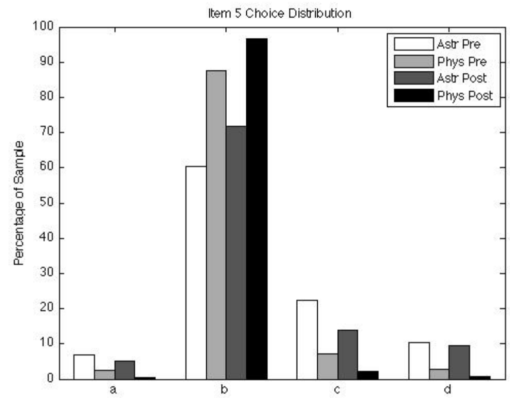
Because some physics students experience the same difficulties as astronomy students, most NGCI items are of appropriate CTT difficulty and discrimination for the physics group. However, Item 5 is the most notable exception, since it has only a weak ability to evaluate physics students’ knowledge and learning. Why do physics students score *so much higher* than astronomy students on Item 5? One hypothesis is that physics courses generally include many more problems like Item 5, such as in kinematics and dynamics lessons, so, rather than having such a far superior understanding of the direction of the gravitational force, physics students’ far superior performance on Item 5 may be due to their familiarity with the context of the problem. If this is indeed the case, it could be an additional explanation for why Item 5 is a poor discriminator of physics student ability.

Table 20: The distribution of student responses to NGCI Items 5, 12, and 26 in the Directionality domain show that Phys 101 students have use proportional reasoning more often than Astro 101 students.

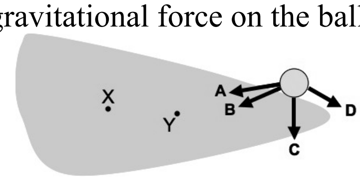
5. A baseball is thrown at an angle so that it follows the dotted path. At the position shown, what is the direction of the gravitational force on the ball?

a. A
b. B
c. C
d. D

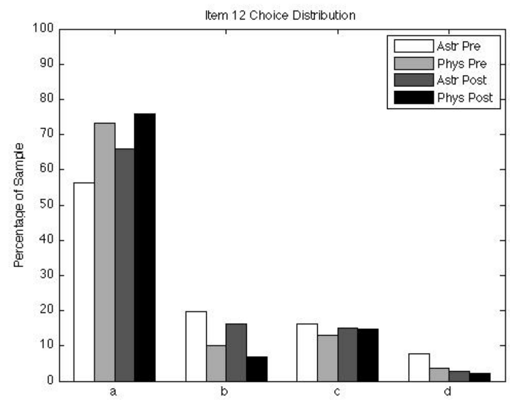




12. For the asteroid shown below, X represents the center of mass and Y represents the geometric center. Which arrow best represents the direction of the gravitational force on the ball?

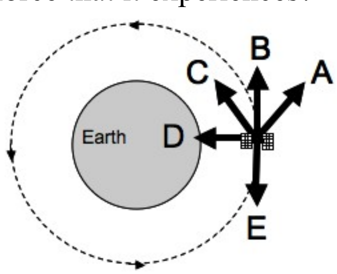


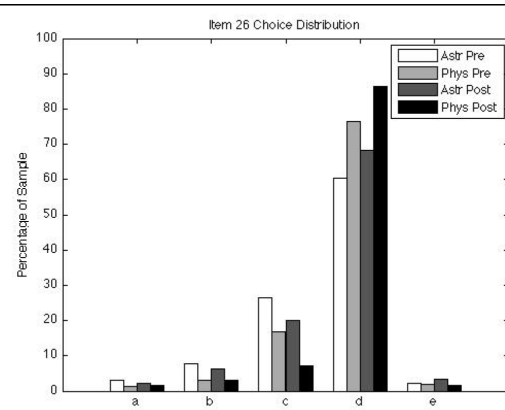
a. A, because it points to X.
b. B, because it points to Y.
c. C, because it points beneath the ball.
d. D, because it points down the slope.



26. An Earth-orbiting satellite is shown at right. Which arrow, if any, best represents the direction of the gravitational force that it experiences?

a. A
b. B
c. C
d. D
e. E





Physics students' familiarity with problems that feature the motion of objects on the surface of Earth may also influence their approach to Item 12, which asks about the direction of the gravitational force for a ball on the surface of an irregularly shaped asteroid. Again, physics students outperform astronomy students on this item, but the most chosen distractor choice for the physics sample is the arrow that points directly beneath the ball (choice 'c'), whereas for the astronomy sample it is an arrow that points to the geometric center (choice 'b'). Again, it is suspected that in working through the physics curriculum of kinematics and dynamics, physics students have much greater familiarity with motion near the surface of a planet (or asteroid). Therefore, in regards to Item 12, physics students may tend to assume that the gravitational force is "perpendicular to the surface" (choice 'b' or 'c,' depending on one's interpretation), rather than toward the center of mass. This "perpendicular to the surface" idea is supported by Interviews F12-04, F12-10, and F12-11 (APPENDIX B and APPENDIX S).

It is concluded that physics students have a superior understanding of the direction of the gravitational force compared to astronomy students – they answered NGCI direction items correctly more often that astronomy students did, Items 5 and 26 are too easy for them, and they use superposition more often (see Item 10 discussed in the Differences in the Force Law domain below), but, there are some important caveats to this statement. First, even though Items 5 and 26 have CTT difficulty values below 0.20, the physics students who answered these items incorrectly display the same types of misconceptions in regards to the Boundary model for the direction of gravity for objects on Earth compared to those in space, and the idea that the direction of gravity is related to the direction of motion. And, second, there is some indication in responses to Item 12 that physics students rely much more on their familiarity with surface problems, reasoning differentially more often that the direction of force is perpendicular to the surface.

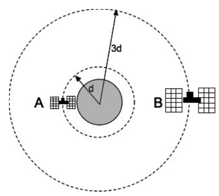
Differences in the Force Law Domain

Another interesting trend observed in the CTT item parameters occurs for items that require proportional reasoning with *both* mass and distance within the gravitational force equation. Items 7, 10, and 16 have somewhat similar CTT difficulty parameters for the astronomy and physics groups (although the physics values are still consistently higher), but much higher discrimination parameters for the physics group. Differences between Phys 101 and Astro 101 student response patterns for these items are shown in Table 21. Item 7 gives only three choices for a comparison of the strength of the gravitational force between Earth and two satellites, one of which is three times more massive and three times farther away than the other. From the response pattern shown in Table 21, one can see that students are almost equally split between the correct answer choice ‘a,’ that the closer satellite experiences the stronger force, and choice ‘c,’ that the two satellites experience the same force. Student interviews suggest that students who choose ‘c’ typically reason that the effects of increasing mass and distance by the same factor cancel out, i.e. there is a one-to-one ratio between mass and distance in the Force Law equation (recall Dostal 2005). However, not all students choosing ‘a’ reason with the correct proportional relationship. Some students could answer correctly by using a simple p-prim such as “closer means stronger.” Here, it is hypothesized that astronomy students are more likely to use this p-prim and that physics students are more likely to use a proportional relationship, so physics students choosing the correct answer are more likely to do so because they have a better overall understanding of Newtonian gravity, making this question more discriminating for the physics sample.

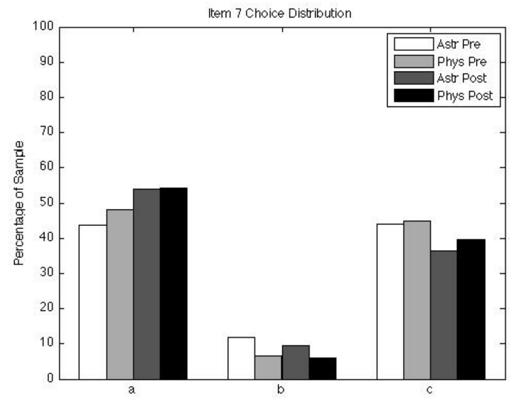
This hypothesis is also supported by the distribution of student responses to Items 10 and 16 (Table 21). Item 10, which primarily probes student ideas in the Directionality domain, requires proficiency in the Force Law domain as well. It asks about the direction of the gravitational force between one planet and two others, one of which is two times more massive and twice as far as other. Here, students have five answer choices, and again, the physics students cluster more often near choices that represent some proportional relationship (choices 'b,' 'c,' and 'd'), whereas astronomy students show differentially higher percentages choosing the extreme choices, 'a' and 'e,' that correspond to reasoning that an object can experience gravitational force from only one other object, which is either the larger object, (i.e. 'larger means stronger'), or the closer object (i.e. 'closer means stronger'), respectively. Finally, Item 16 asks students to compare the gravitational force between a small planet and another of equal mass and one that is four times as massive but twice as far as the first. Again, one can see a much higher percentage of post-instruction physics students choosing the correct answer that corresponds to more sophisticated proportional reasoning, while astronomy students' responses are more evenly distributed.

Table 21: The distribution of student responses to NGCI Items 7, 10, and 16 in the Force Law domain show that post-instruction Phys 101 use proportional reasoning differentially more often than Astro 101 students.

7. Satellite B is *three* times more massive than A, but orbiting the planet at *three* times the distance. Compare the force of gravity between each satellite and the planet.

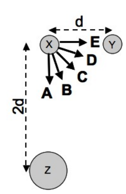


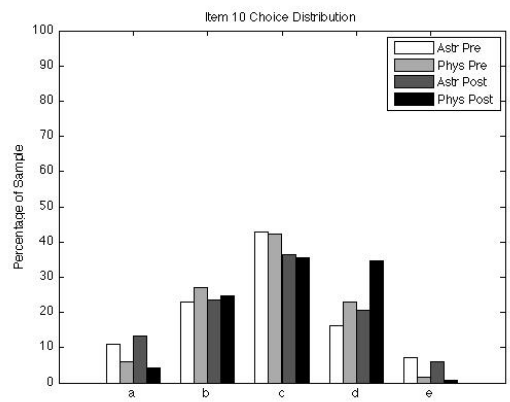
a. A experiences a stronger force than B.
b. B experiences a stronger force than A.
c. A and B experience the same force.



10. 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?

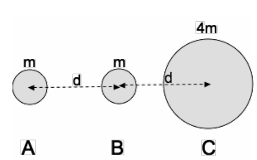
a. A
b. B
c. C
d. D
e. E

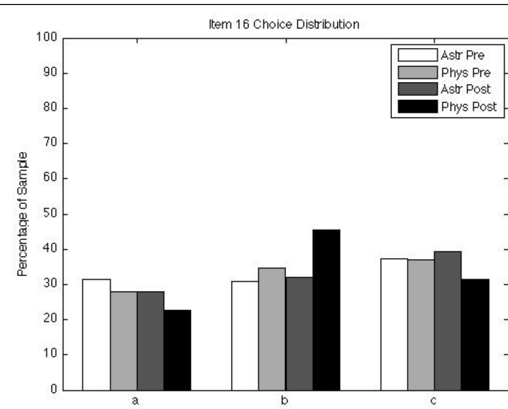




16. The following diagram shows two smaller planets of mass *m* and one larger planet of mass *4m*, each separated by distance *d*. The gravitational force *by Planet C on Planet A* is ______ that *by Planet B on Planet A*.

a. is less than
b. is equal to
c. is greater than



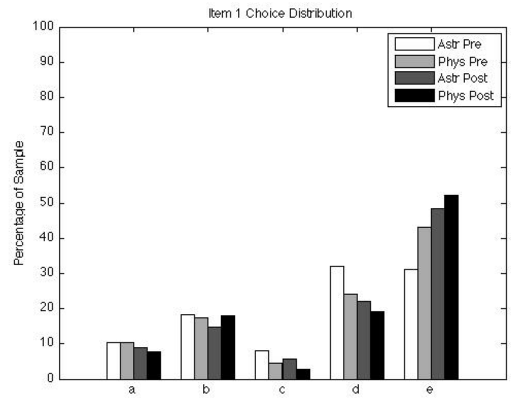


Differences in the Independence
of Other Forces Domain

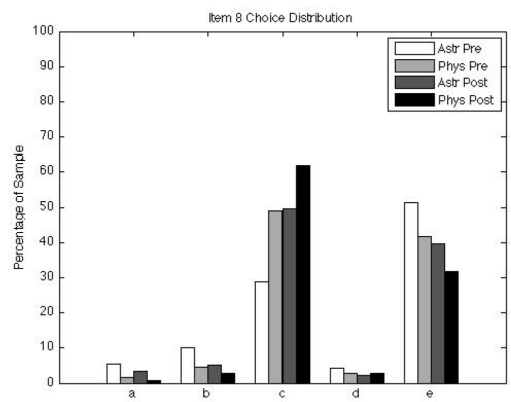
Items in the Independence of Other Forces domain have somewhat similar CTT difficulty and discrimination values for both the astronomy and physics samples, but a closer inspection of students' choice distributions for Items 1 and 8 (Table 22) reveals that some distractors function differently for the two groups. For instance, Item 1 asks which factors (if any) other than gravity (air pressure, forces from spinning, magnetism, or some combination of these) hold us to Earth's surface. Pre-instruction, choice 'd', "More than one of the above factors," was the most effective distractor for both populations; however, post-instruction, this remained the most effective distractor for the astronomy population only. The most chosen post-instruction distractor for the physics sample was 'b', "Forces from Earth's spinning motion." A similar trend is observed in the response pattern of Item 8 (Table 22), which asks why Earth exerts a gravitational force on objects on its surface. Again, pre-instruction, the distractor choices function similarly for astronomy and physics students. Post-instruction, all distractor choices were less appealing to astronomy students, but about the same proportion of physics students chose 'd', "It rotates" as pre-instruction physics students. Again, physics students choose this distractor proportionally more often than astronomy students. While these differences are small, they seem to allude to a real bias of physics students to confound gravity with rotation.

Table 22: The distribtuion of student responses to NGCI Items 1, 8 and 9 in the Other Forces domain show that post-instruction Phys 101 students are drawn differentially more often than Astro 101 students to the "rotation" distractor choice.

1. Besides the force of gravity, which of the following factors hold(s) us to Earth's surface?
   a. Air pressure from Earth's atmosphere.
   b. Forces from Earth's spinning motion.
   c. Magnetism from Earth's magnetic field.
   d. More than one of the above factors.
   e. No other significant factors; only gravity.

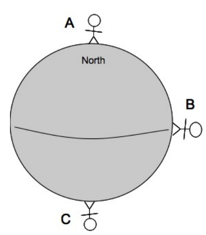


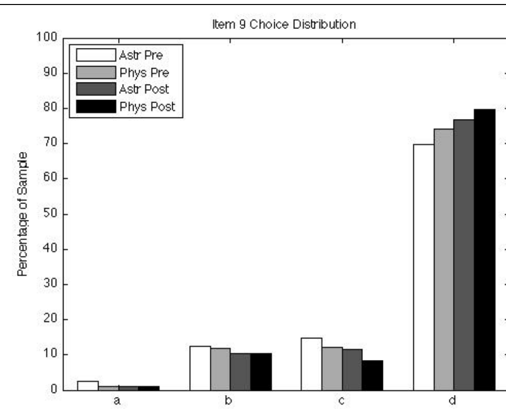
8. Why does Earth exert a gravitational force on objects on its surface?
   a. It has an atmosphere
   b. It has a magnetic field.
   c. It has mass.
   d. It rotates.
   e. More than one of these.



9. Assume Earth is perfectly round and that a person walks from the North Pole to the South Pole. At which position does the person experience a stronger force of gravity from Earth?

   a. A only
   b. B only
   c. Both positions A and C
   d. Each point is the same





Why did the functioning of the rotation distractor choices for Items 1 and 8 change pre-instruction to post-instruction for the physics sample but not the astronomy sample? Interviews suggest that astronomy students choose the distractor about rotation when reasoning with commonsense experience that moving in circular motion causes one to 'feel' lighter. However, through explicit lessons and homework problems on circular motion, physics students generally encounter rotational forces more often. They must confront the apparent contradiction with daily experience that centripetal force is an inward force. However, they may not understand the nuances of the effects of reference frame and could conclude that, for people on a rotating Earth, centripetal force is a downward force in addition to gravity. Additionally, weight forces frequently appear on free-body diagrams and are used in calculations of centripetal force in physics curricula. Therefore, centripetal motion is a topic that is incorporated into physics students' mental landscapes alongside weight (and, therefore, gravitational force), making some physics students more likely to choose the rotation distractor.

In this context, physics student responses to Item 9 (also shown in Table 22) might also make more sense. Item 9 asks about the force of gravity a person at different latitudes on Earth, and (as seen in Figure 4) it is the least difficult item for the astronomy students but not for the physics students. Student response patterns show that physics students are drawn more to the distractor that a person at the equator would experience a stronger force of gravity (choice 'b'). While none of the physics students who chose option 'b' volunteered to be interviewed, it is speculated that the reasoning for choosing this distractor is the same as it is for the astronomy students, i.e. that a person at the equator would be at the fastest spinning part of Earth. Therefore, physics students' bias toward compounding gravity with rotation could make Item 9 more difficult than some other items on the NGCI.

This bias appears to be specific to only the rotation distractors in the Mixing of Forces domain. While, like astronomy students, some physics students are drawn to magnetic and atmospheric distractors, they do not exhibit any obvious preference for these. The proportions of students choosing distractors that correspond to magnetic forces (choice 'c' in Item 1 and choice 'b' in Item 8) and atmospheric forces (choice 'a' for both items) decrease approximately equally pre- to post-instruction for both groups. Additionally, Item 15, which explicitly asks about how gravity and magnetism are related, does not show a difference in response patterns. Therefore, it is concluded that physics and astronomy students share difficulties within the Mixing of Forces domain in terms of magnetic and atmospheric forces, but physics students experience slightly more difficulties with rotational forces.

Differences in the Threshold Domain

After Item 5 (discussed above in the Differences in Directionality domain section), the next *five* most differentially functioning physics and astronomy CTT difficulty values fall within the Threshold domain (refer to the expert categorization in Table 16). Again, these were items whose physics and astronomy CTT values were farthest apart in Figure 4 – Items 2, 13, 20, 21, and 23. Item 2 asks about the gravitational force between two astronauts in space, Item 13 asks what is required for the gravitational force between two objects to be zero, Item 20 asks what would happen to a pen released from an astronaut's hand on the Moon, Item 21 asks which objects (Sun, Moon, Person, Mars) experience a gravitational force from Earth, and Item 23 asks how the gravitational force between Earth and a rocket would diminish as the rocket leaves Earth. Difficulties that students experience with these items are conceptually similar: APPENDIX Q shows that the relative strengths of distractor choices for all of these questions were the same for both physics and astronomy students, these items had equally good discrimination values for both physics and astronomy students, and interviews suggest that students who answer incorrectly reason with the idea that there is a minimum mass and/or maximum distance or atmospheric threshold for which gravity "stops." However, these questions were all *much* easier (i.e. much lower CTT difficulty values) for the physics students both pre- and post-instruction. It is concluded that the NGCI items in the Threshold domain probe similar misconceptions for physics and astronomy students, but the physics students have a truly more sophisticated understanding of the universality of gravity in limiting cases, with the Threshold domain being the area where physics students excel most over astronomy students.

Discussion

The pilot testing of the NGCI with Phys 101 students showed that the NGCI is a reliable and valid instrument for probing Phys 101 students' understanding of Newtonian gravity, addressing the first research question posed at the beginning of this chapter. To address the second research question, the previous sections show that the NGCI can tease apart many differences in Astro 101 and Phys 101 students' understanding of Newtonian gravity. First, comparing the pilot site data in Table 5 and Table 17 shows physics students have generally higher pre-instruction and post-instruction NGCI scores, indicating that they have a superior overall understanding of Newtonian gravity. Second, combined with student response patterns, comparing the physics and astronomy CTT statistics shows important difference between physics and astronomy students in each of the four NGCI conceptual domains. Together with student interviews, these data indicate that physics students generally think more mathematically (i.e. they more often wrote equations during interviews and chose distractors corresponding to proportional reasoning) and scientifically (i.e. they adhered more to the "universality" of gravity and appeared to rely more on their experiences in the classroom rather than their intuition). However, some physics students suffer from the same misconceptions as astronomy students, most notably in regard to the Mixing of Forces model, where rotational motion is more incorporated into their mental landscapes. Though far less prevalent, some physics students still chose distractors corresponding to: the Mixing of Forces model with magnetism and atmospheric forces, the Boundary Model with an atmospheric or orbital boundary, and a misapplication of proportional reasoning in the gravitational force equation. Interpretive arguments related to these misconceptions are explored in the sections above within each of the four NGCI conceptual domains and can be summarized as follows:

- Directionality: Physics students have a greater understanding of the directionality of gravity, but they seem to exhibit a bias for the "perpendicular to the surface" idea.
- Force Law: Physics students use proportional reasoning with mass and distance more frequently than astronomy students.

- Independence of Other Forces: While physics and astronomy students both have misconceptions related to the Mixing of Forces mental model, physics students experience slightly fewer difficulties related to magnetism and atmospheric forces and slightly greater difficulties in regard to rotational forces.
- Threshold: Physics students excel most above astronomy students in reasoning about gravity in cases that test the universality of gravity for objects that have small masses, are separated by large distances, and regardless of perceived atmospheric or orbital "boundaries."

Furthermore, this analysis provides additional evidence to argue for the validity of the NGCI as a robust tool to measure student understanding of gravity. While the Classical Test Theory statistics alone confirm the NGCI's applicability to both Astro 101 and Phys 101 populations, comparing CTT parameters and exploring why student response patterns may differ between these two populations shows that the NGCI is highly sensitive to the most prevalent student difficulties. All four conceptual domains of the NGCI are uniquely sensitive to how these difficulties manifest between Astro 101 and Phys 101 students and curricula. No other research-validated instrument has been shown to make such robust comparisons between these two groups, making the NGCI especially unique and powerful. The remainder of this dissertation further expands the NGCI's capabilities by (1) *calibrating* the items independently of the sample population and (2) *measuring* Phys 101 and Astro 101 students' learning of Newtonian gravity independently of the NGCI items and controlling for student and classroom characteristics.

CHAPTER SIX

ITEM RESPONSE THEORY CALIBRATION

As discussed in Chapter 1, the concept of measurement of student understanding depends on a standard measurement device, or ruler, just like any other physical measurement instrument (Wilson 2005). The NGCI has been shown to measure student understanding of Newtonian gravity only in the sense of ranking (Planinic, Ivanjek, & Susac 2010, Wallace & Bailey 2010), i.e. a higher score is associated with a greater understanding of gravity. But, in comparing Phys 101 and Astro 101 students, one wants to know: *How much* do *students' understandings* differ (1) pre-instruction, (2) post-instruction, and (3) as the result of instruction (i.e. how much do they learn in their respective classes?). A simple comparison of total scores is not sufficient for answering any of these questions, because the score is an *outcome* of a student's understanding upon interacting with a test, but not her understanding itself (Hambleton & Jones 1993, Ding & Beichner 2009, Ivanjek, Planinic, & Susac 2010). To adequately calibrate NGCI items for all Astro 101 and Phys 101 students, a method that can calibrate item characteristics such that they are independent of the sample population is needed. Similarly, to investigate the question of *how much* students' understandings differ, student abilities need to be measured along a linear construct that is independent of the instrument. While Classical Test Theory proved straightforward and useful in the development of the NGCI in Chapter 4 and for comparing Astro 101 and Phys 101 students in Chapter 5, it has some important shortcomings. Namely, CTT statistics are highly sample dependent

(Hambleton & Jones 1993, Fan 1998, Wallace & Bailey 2010), so the CTT item difficulty and discrimination parameters presented above apply only to the very specific students that participated in this study and cannot be extended to other samples of Astro 101 and Phys 101 students. To calibrate and extend the applicability of Concept Inventories beyond these shortcomings of Classical Test Theory statistics, the PER and AER communities have adopted the logistic functions of Item Response Theory (IRT) (for example, Lee et al 2008, Planinic, Ivanjek, & Susac 2010, Wang & Bao 2010, and Wallace & Bailey 2010). The basic tenants of IRT and the application of IRT to the NGCI are discussed below.

Item Response Theory

Item Response Theory (IRT) assumes that a dichotomously scored survey instrument like the NGCI measures student ability along a unidimensional continuum, i.e. the construct of Newtonian gravity. Mapped along this continuum are different levels of understanding, or latent intrinsic, though changeable (Wallace & Bailey 2010), ability levels, which increase linearly from a low ability to a high ability (Wilson 2005). Item characteristics are also modeled as intrinsic properties of the items, such that even if a researcher happens to have a sample of students of very low or very high ability students, the same item parameters would apply. IRT then provides a *probability* estimate of a student of a given ability answering a given item correctly (Hambleton & Swaminathan 1985, Hambleton & Jones 1993, Fan 1998), thus accounting for some inherent uncertainty in person-test interaction. Pedagogically, this complements the cognitive research on how students' mental landscapes, or latent abilities, manifest in response to cues and contextual features of a problem, as discussed in Chapter 1. Furthermore, unlike CTT, IRT provides a way to determine how well the statistical model fits the data (Hambleton & Jones 1993, Wallace & Bailey 2010). The number of parameters to be estimated, the number of students in the calibration sample, and the degree to which the instrument measures a single construct all influence model fit (Harris 1989), so each IRT item parameter and student ability estimate has a corresponding uncertainty value. Consequently, *when the IRT model fits the data, student abilities are item-independent and item characteristics are population-independent*, leading to the so-called "parameter invariance" of IRT models that make them so appealing despite their relatively strict requirements and cumbersome calculations (Hambleton & Swaminathan 1985, Crocker & Algina 1986, Hambleton & Jones 1993, Fan 1998, Rupp & Zumbo 2006, Wallace & Bailey 2010, and references therein).

<u>IRT Logistic Functions</u>

What does an IRT model look like? IRT characterizes a student *p*'s correct response $X_{pi} = 1$ to a particular item *i* as a probability function of the student's latent ability, θ_p, in the context of item features (such as difficulty parameter, b_i, discrimination parameter, a_i, and guessing parameter, c_i) via a logistic function such as the Three-Parameter Logistic (3PL) function shown below (Crocker & Algina 1986, Yen & Fitzpatrick 2006):

$$P\left(X_{pi} = 1 | \theta_p, a_i, b_i\right) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_p - b_i)}}$$

Note that this applies for any student, so the independent variable is the range of student abilities, θ's, which are measured in "logits," or log-odds units of answering the item correctly and scaled to have a mean of zero and standard deviation of one (Crocker & Algina 1986, and see additional details about this process in APPENDIX T). For each item, the function $P(\theta)$ is represented graphically with what is known as an Item Characteristic Curve, or ICC (Crocker & Algina 1986, Harris 1989). The graph in Figure 6 below shows ICCs for three example items. The S-shaped curve of an ICC shows that students at low ability levels have a low probability of answering an item correctly while students at high ability levels have a high probability. Item parameters change the shape and position of the curve (Crocker & Algina 1986, Harris 1989). The discrimination parameter, a, modifies the slope at the midpoint of the curve, with higher values creating a steeper slope and indicating that the item has strong discriminatory power. In this example, Item III is the most discriminating item. The difficulty parameter, *b*, is the location of the midpoint of the curve, with higher values shifting the curve to the right and indicating a more difficult item. IRT difficulty values, b's, are measured in logits on the same scale as student ability values, θ's (Harris 1989, Hambleton & Jones 1993, Fan 1998). Item I in the example is the most difficult, because even students of higher ability have a low probability of answering the item correctly. One will note that the IRT difficulty and discrimination parameters are different than those in CTT (Wallace & Bailey 2010). Finally, the guessing parameter, *c*, represents the lower asymptote of the ICC. Item II has the highest guessing parameter, indicating that even students at low ability have a high probability of answering the item correctly just by chance.

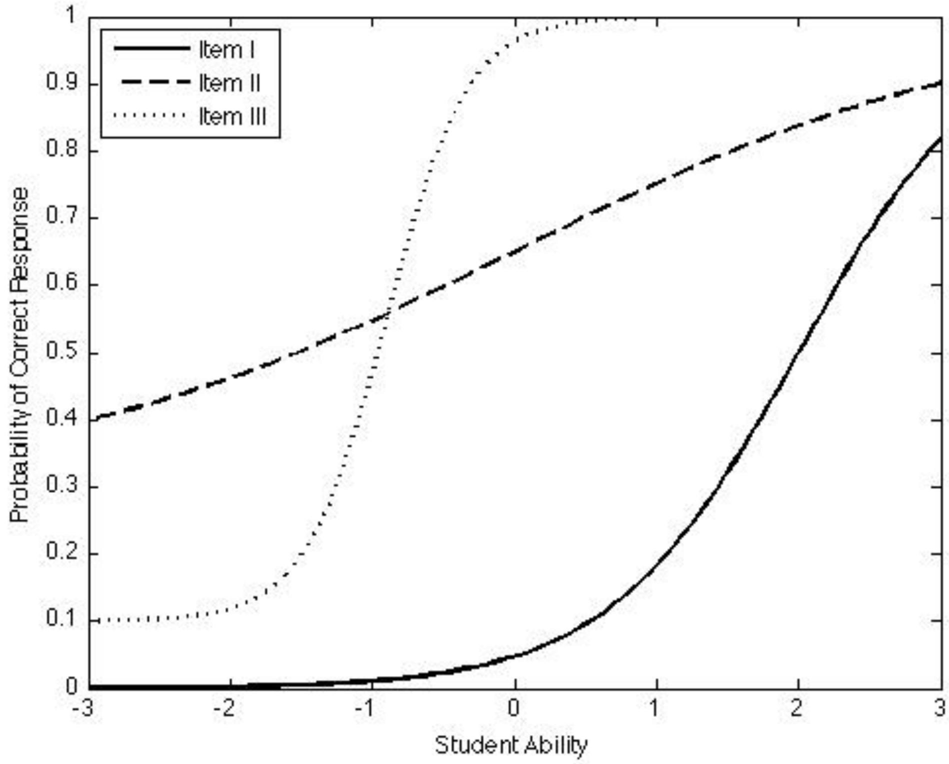


Figure 6: Example ICCs for three items: Item I (a = 1.5, b = 2.0, c = 0.0), Item II (a = 0.6, b = 0.0, c = 0.3), and Item III (a = 3.5, b = -0.9, c = 0.1).

Other IRT logistic functions include the Two-Parameter Logistic (2PL) Model, which is the same as the 3PL Model with the guessing parameter, *c*, set to zero, and the Rasch Model, which sets *c* to zero and the discrimination parameter, *a*, to one (or some other constant, i.e. it assumes that all items have the same discriminatory power). A researcher chooses which IRT model to use based on the purposes of the research questions and the degree to which the model interpretations match the context of the problem (Andrich 2004). With fewer parameters to fit, the Rasch Model is the more rigid model and is typically chosen when one is interested in modifying items to fit the population, while the 3PL model generally fits the data better and can therefore provide wider interpretations of the population (Crocker & Algina 1986, Harris 1989). As discussed below, the 2PL model was chosen for calibration of the NGCI data.

<u>Using the 2PL Model</u>

This research project used a four-group 2PL IRT model for analysis of the NGCI, because (1) the 925 pre-instruction astronomy students, 1,392 pre-instruction physics students, 743 post-instruction astronomy students, and 929 post-instruction physics students constitute a large enough sample size (n=3,989) for using an IRT model to begin with (Wang & Bao 2010), (2) the CTT quantitative analysis and validation process discussed in Chapters 4 and 5 show that the NGCI items are appropriate for the Astro 101 and Phys 101 populations and vary in discriminatory power, making the Rasch model unnecessary, and (3) the distribution of student response patterns (APPENDIX Q) and student interviews (APPENDIX C, APPENDIX R, and APPENDIX S) suggest that students are very rarely outright guessing on items[9], making the 3PL model an unnecessary model. Indeed, as discussed in Chapters 3 and 4, the NGCI distractors were carefully designed to appeal to students' known reasoning patterns. Furthermore, using the IRT difficulty and discrimination parameters provides context for comparisons with CTT values, even though they are not on the same scale (see the NGCI Calibration section).

In fitting a 2PL model to dichotomously scored student response data, one must solve for a variety of unknowns. There are two parameter estimates (*a* and *b*) for each of

[9] While sometimes interviewees claimed to be guessing, they often had *some* reason for choosing one answer over another, with "guessing" seeming to merely reflect their lack of confidence in choosing that answer.

the N=26 items, and there is a unique ability estimate for each of the n=3989 student responses in the four-group sample These 2N + n values can be estimated via a variety of procedures accomplished with computer programs that maximize the likelihood of observed response patterns (Harris 1989, Wallace & Bailey 2010). For the four-group 2PL analysis of the NGCI, the software BILOG-MG (Zimowski et al 1996) was used to implement a marginal maximum likelihood (MML) estimation of NGCI item parameters and an expected a posteriori (EAP) estimation of student abilities. One may refer to APPENDIX T for an overview of these estimation procedures and references therein for more fundamental treatments.

Testing IRT Assumptions

The two assumptions of IRT are unidimensionality and local independence (Crocker & Algina 1986, Wallace & Bailey 2010). First, if an instrument is unidimensional, there is a single construct being measured and ability estimates fall along a single linear scale (Wilson 2005). Accordingly, each item probes proficiency along the underlying construct only (i.e. understanding of Newtonian gravity for the NGCI) and not along any other trait (factors that the instrument development process attempted to minimize such as familiarity with jargon, understanding of planet formation, ability to read graphs, etc.). While instruments are never perfectly unidimensional, they should come as close to unidimensionality as possible for using IRT (Hambleton & Jones 1985). Second, if the items of an instrument are locally independent, then a student's response on one item does not influence her response to another item. Techniques for evaluating the degree to which an instrument meets these criteria are discussed below.

A variety of techniques have been recommended for assessing unidimensionality. Following the precedent of Planinic, Ivanjek, and Susac (2010), one can check the Pearson correlation between students' ability estimates and their responses to items. According to their paper, one wants all items to point in the same direction, i.e. all correlations should be positive, and "the size of the correlations can indicate which items contribute more to the construct and which contribute less." From Table 23, one can see *all positive correlations*, with 25 items showing a medium-strength correlation (between 0.30 and 0.80) and only one item (Item 7) showing a low correlation, indicating that all items contribute to the underlying construct.

Table 23: NGCI items' Pearson correlations between students' IRT abilities and their responses to each item.

Item	Correlation	Item	Correlation
1	0.43	14	0.44
2	0.59	15	0.56
3	0.35	16	0.31
4	0.55	17	0.39
5	0.54	18	0.36
6	0.40	19	0.48
7	0.27	20	0.60
8	0.72	21	0.50
9	0.33	22	0.62
10	0.40	23	0.51
11	0.38	24	0.38
12	0.49	25	0.32
13	0.47	26	0.51

Furthermore, Hambleton and Swaminathan (1985) and Wang and Bao (2010) recommend performing an eigenvalue analysis with correlations between results of different questions via principle component analysis. If the primary eigenvalue is much larger than the other eigenvalues (Wang and Bao 2010 recommend an order of magnitude, but accept a factor of approximately 4.6 for their study), then that factor explains most of the observed variance and the instrument can be deemed unidimensional. This was performed for the NGCI data using the IBM SPSS statistical software, and as shown in the “scree plot” of eigenvalues (Figure 7), the primary eigenvalue is 3.6 times larger than all other eigenvalues. Taken together, the Pearson correlation values and the eigenvalue analysis imply that the NGCI is sufficiently unidimensional.

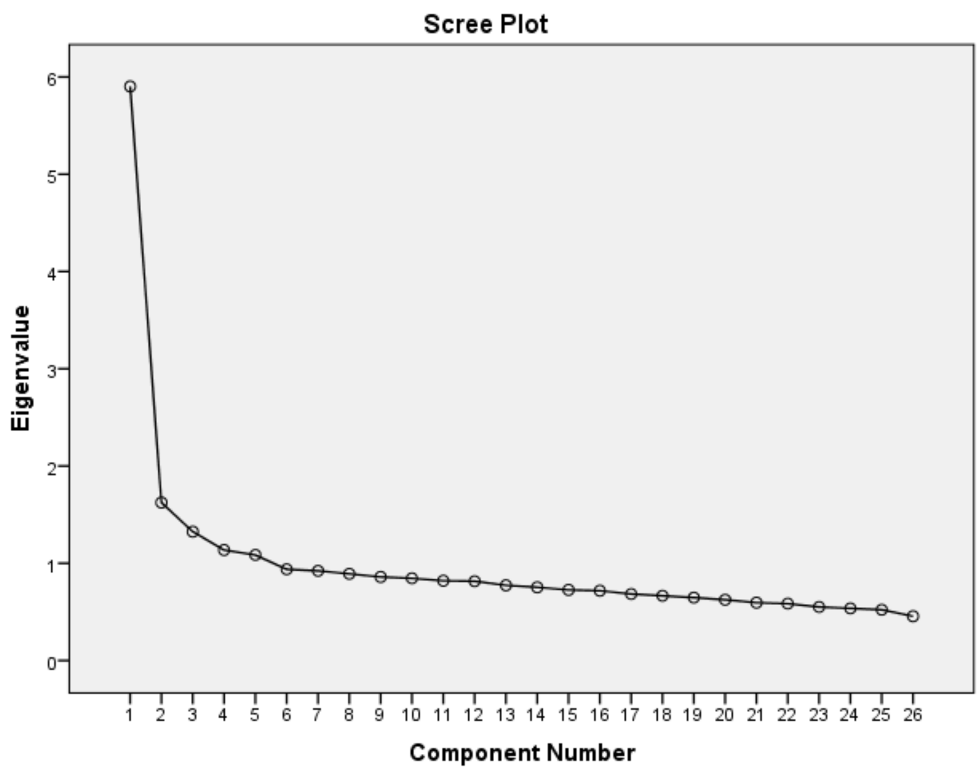


Figure 7: Scree plot of factors (horizontal axis) and their eigenvalues (vertical axis) shows that there is one principle factor being measured by the NGCI and that the instrument is unidimensional.

Local independence assumes that a student at any fixed ability level does not answer one item conditional on her answer to another item. While Wang and Bao (2010) argue that if an instrument is unidimensional the items are automatically locally independent, the local dependence of NGCI items was also evaluated with Yen's (1984) Q3 statistic as in Wallace and Bailey (2010). Yen's Q3 provides a way to test dependence between items by correlating the residual differences between expected and observed responses (Wallace & Bailey 2010). If the items are locally independent, these correlations should be small, and Yen and Fitzpatrick (2006) recommend that all item pair correlations should be <|0.20|.

After performing a 2PL analysis of the NGCI data, Yen's Q3 showed residual correlations greater than |0.20| for four item pairs: Items 7 and 10 (0.25), Items 11 and 14 (0.29), Items 11 and 17 (0.24), and Items 14 and 17 (0.25). Items 10, 14, and 17 were eliminated from the analysis, leaving the items that appear earlier in the NGCI, under the assumption that students' answers to earlier (i.e. lower number) questions would be more authentic. After eliminating these items, Items 4 and 7 also showed a residual correlation (0.20), so Item 7 was also eliminated from the analysis. Yen's Q3 values for the remaining 22 items are all below |0.20|, as shown in APPENDIX U. To ensure a locally independent data set, student responses to these 22 items only were used for the final 2PL IRT analysis of the NGCI.

NGCI 2PL IRT Results

Table 24 shows the 2PL item parameters, discrimination parameter, *a*, and difficulty parameter, *b*, with accompanying standard errors calculated with BILOG-MG (Zimowski et al 1996). Infit and Outfit values are also provided to assess how well the 2PL model fits the data, as discussed below. Even though the 22 NGCI items satisfy both IRT assumptions (unidimensionality and local independence), meaningful interpretations of these results can be made only if the 2PL model fits the data (Wallace & Bailey 2010).

Table 24: 2PL IRT item results for the 22-item subset of NGCI data. Infit and outfit values below 0.75 and above 1.33 are bolded and italicized.

Item	a	a's SE	b	b's SE	Infit	Outfit
1	0.68	0.03	1.24	0.06	0.95	0.94
2	1.19	0.05	0.68	0.03	0.87	0.84
3	0.57	0.03	-0.03	0.07	0.94	0.92
4	0.99	0.04	1.43	0.04	0.92	0.92
5	1.94	0.11	-0.57	0.03	0.87	***0.61***
6	0.70	0.04	-0.18	0.06	0.92	0.89
8	1.79	0.07	0.79	0.03	0.78	***0.70***
9	0.62	0.04	-1.18	0.11	0.93	0.89
11	0.69	0.04	-0.63	0.08	0.91	0.88
12	0.98	0.05	-0.26	0.05	0.89	0.82
13	0.79	0.04	0.72	0.05	0.92	0.90
15	1.06	0.04	0.68	0.04	0.88	0.85
16	0.51	0.03	2.10	0.10	1.00	1.02
18	0.58	0.03	0.03	0.07	0.94	0.92
19	1.04	0.05	-0.54	0.05	0.87	0.81
20	1.57	0.07	-0.13	0.03	0.85	***0.68***
21	1.04	0.05	-0.03	0.04	0.91	0.82
22	1.62	0.07	-0.02	0.03	0.84	***0.67***
23	0.99	0.05	0.09	0.04	0.89	0.83
24	0.62	0.03	0.41	0.06	0.94	0.93
25	0.61	0.03	3.66	0.15	1.10	1.11
26	1.19	0.06	-0.41	0.04	0.88	0.77

Assessing Model Fit

While there is no single method to determine model fit (Harris 1989, Yen & Fitzpatrick 2006), one method by which it can be assessed is by calculating the squared residuals between what is observed in the data, *X*, and what is expected from the model, P, and standardizing these residuals by dividing by the expected variance ($\sigma^2 = P(1 - P)$ for dichotomously scored items), such as with the infit and outfit statistics in Table 24 (Bond & Fox 2001, Wilson 2005, de Ayala 2009, and Planinic, Ivanjek, & Susac 2010). Infit is the ratio of the sum of the squared residuals compared to the sum of the expected variances across all respondents, *p*:

$$Infit = \frac{\sum_{p=1}^{n}(X - P)^2}{\sum_{p=1}^{n}\sigma^2}$$

Infit values greater than 1.00 indicate that the observed variance is greater than the expected variance (Wilson 2005), and that students whose abilities are near the difficulty of the item are answering in an unexpected way. Outfit values, on the other hand, standardize the residuals for each student by dividing by the expected variance and *then* summing these across all students. It is then necessary to divide by the number of respondents *n* to normalize. Outfit values, therefore, give more weight to the effects of respondents far from the item's sensitivity, so that high outfit values indicate that students' far from what the model predicts are performing in an unexpected way:

$$Outfit = \frac{1}{n}\sum_{p=1}^{n}\frac{(X - P)^2}{\sigma^2}$$

Both infit and outfit values should be close to 1.00, with large infit values considered more problematic than large outfit values (Planinic, Ivanjek, & Susac 2010); however, the range of what is acceptable is debated. Bond and Fox (2001) suggest a conventional range of 0.75 to 1.33, Planinic, Ivanjek, and Susac (2010) used cutoffs of 0.5 and 1.5 for their study, and Wu and Adams (unpublished) argue that these cut-offs should be modified based on sample size, so that 95% of items should have infit and outfit values between $1 \pm 1.9\sqrt{2/n}$, indicating a range of 0.96 to 1.04 for the n = 3,989 students in this study. Table 24 flags items that are in the range of Bond and Fox (2001). All infit values range from 0.78 to 1.10 and four items (Items 5, 8, 20 and 22) are flagged with somewhat low outfit values, indicating that the observed variance for these items is less than expected, i.e. the model is "over fitting" the data. Additionally, both infit and outfit values for Item 25 are higher than 1.00, indicating that it may contribute less than the other items to measuring the underlying construct of Newtonian gravity ability (discussed in the Calibrating the NGCI section below).

As an additional way to visualize model fit, one can plot the Item Characteristics Curve (ICC) of each item with accompanying predicted standard deviation bounds overlaid with observed data (Hambleton & Swaminathan 1985, Harris 1989, and Wallace & Bailey 2010), such as in the plots shown in Table 25. For plotting the observed data points, students were grouped according to ability level into 42 bins of approximately 95 students each. The x- and y-values for the comparison on the ICC plots, then, are the bin's average ability and the proportion of students in each bin who correctly answered the question, respectively. Inspection of these plots shows that the data generally map out the model predictions.

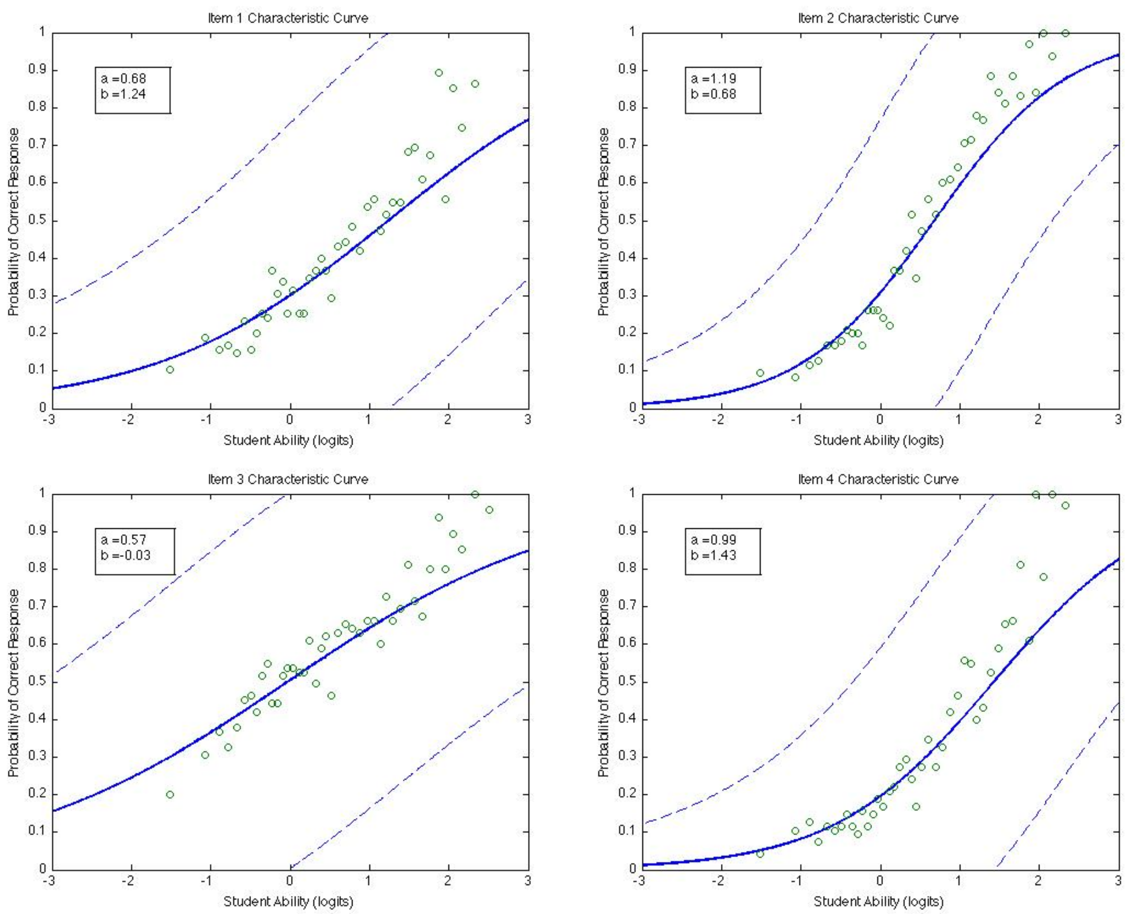
So, does the 2PL IRT model fit the data? Considering the fit statistics and ICC's with overlaid data in context with the purposes of the instrument, it is up to the researchers to make a judgment call (Hambleton & Swaminathan 1985, Harris 1989, and Hambleton & Jones 1993). If the assessment provided by the instrument is used to make strict judgments about a person, stringent fit between model and data may be required. However, since the NGCI measures student understanding, rather than a more stringent trait such as for medical assessment, and since IRT can provide an instructive calibration of the NGCI, it is concluded that the 2PL IRT model adequately fits the four-group, 22-item NGCI data for the purposes of this project.

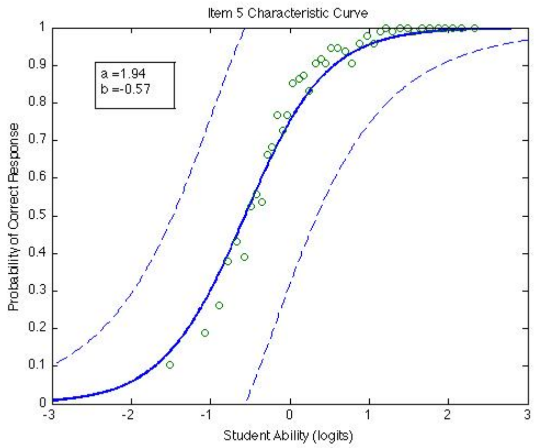
Because the 2PL IRT model adequately fits the four-group NGCI data, it is asserted that the IRT results are *invariant*. This means that the item parameter estimates in Table 24 are taken as *independent of the sample population used to calculate them*, thereby able to measure Newtonian gravity ability in *all* Astro 101 and Phys 101 students (Rupp & Zumbo 2006)[10]. Recall that CTT parameters were different for each of the four groups in the astronomy and physics pre- and post-instruction samples. So, even if the sample population represents a special subset of only low or high ability students from the larger population, the same item parameters would apply (Wallace & Bailey 2010). Furthermore, student ability estimates *represent students' "true" abilities independent of the particular items that were used to measure them*. Again, this offers an advantage over the traditional methods of exploring student understanding that rely on students' total

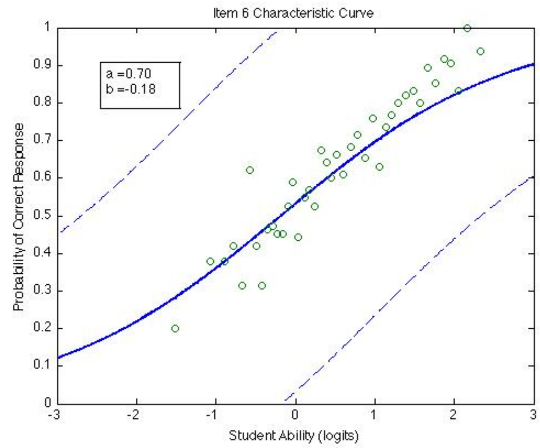
[10] Recall from APPENDIX T that the zero-point and width of the logit scale are anchored so that the mean of the ability distribution is zero and the standard deviation is one, meaning that a linear transformation is needed to translate parameters to the same scale (Crocker & Algina 1986, Rupp & Zumbo 2006).

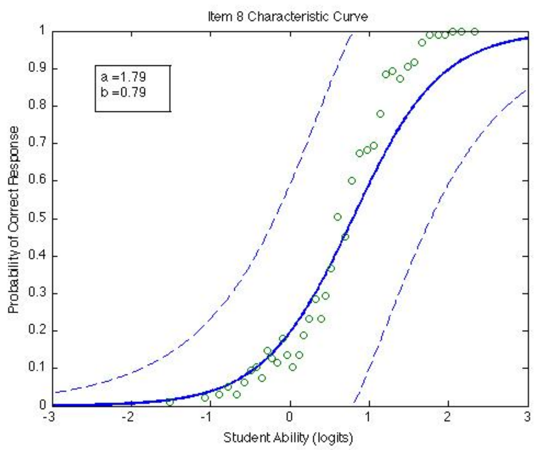
scores. The following sections use the 2PL IRT model results presented above to discuss the *calibration* of the NGCI and the *measurement* of student Newtonian gravity ability.

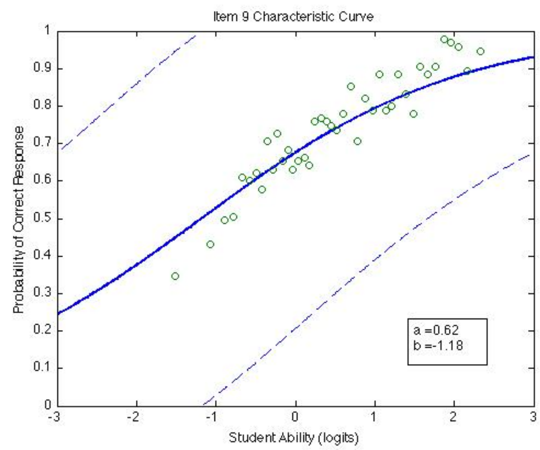
Table 25: The four-group, 22-item NGCI 2PL IRT Item Characteristic Curves. Solid blue line is theoretical curve, dotted lines are expected standard deviation. Green circles are the actual observed percentage of students answering correctly within that interval of ability. 42 bins were used with approximately 95 students per bin.

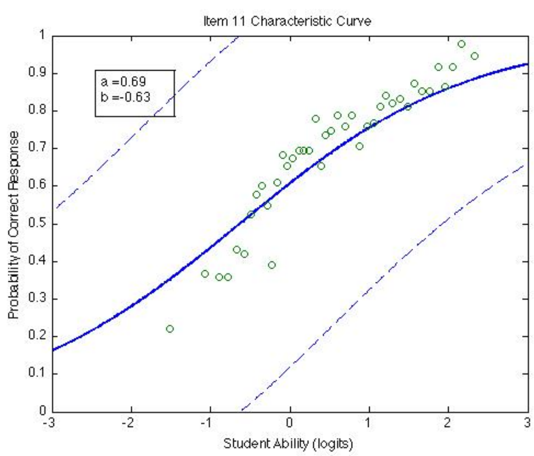


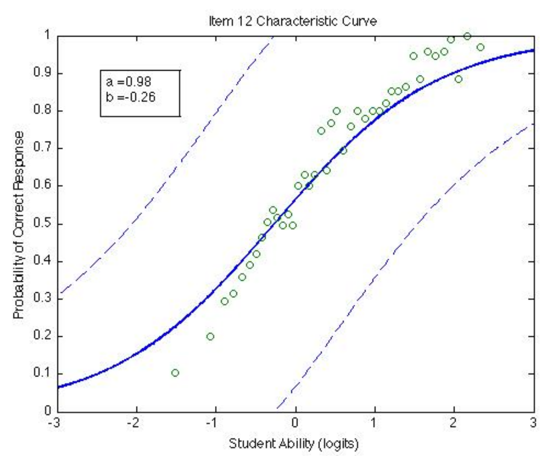

Item 5 Characteristic Curve
a =1.94
b =-0.57
Probability of Correct Response
Student Ability (logits)

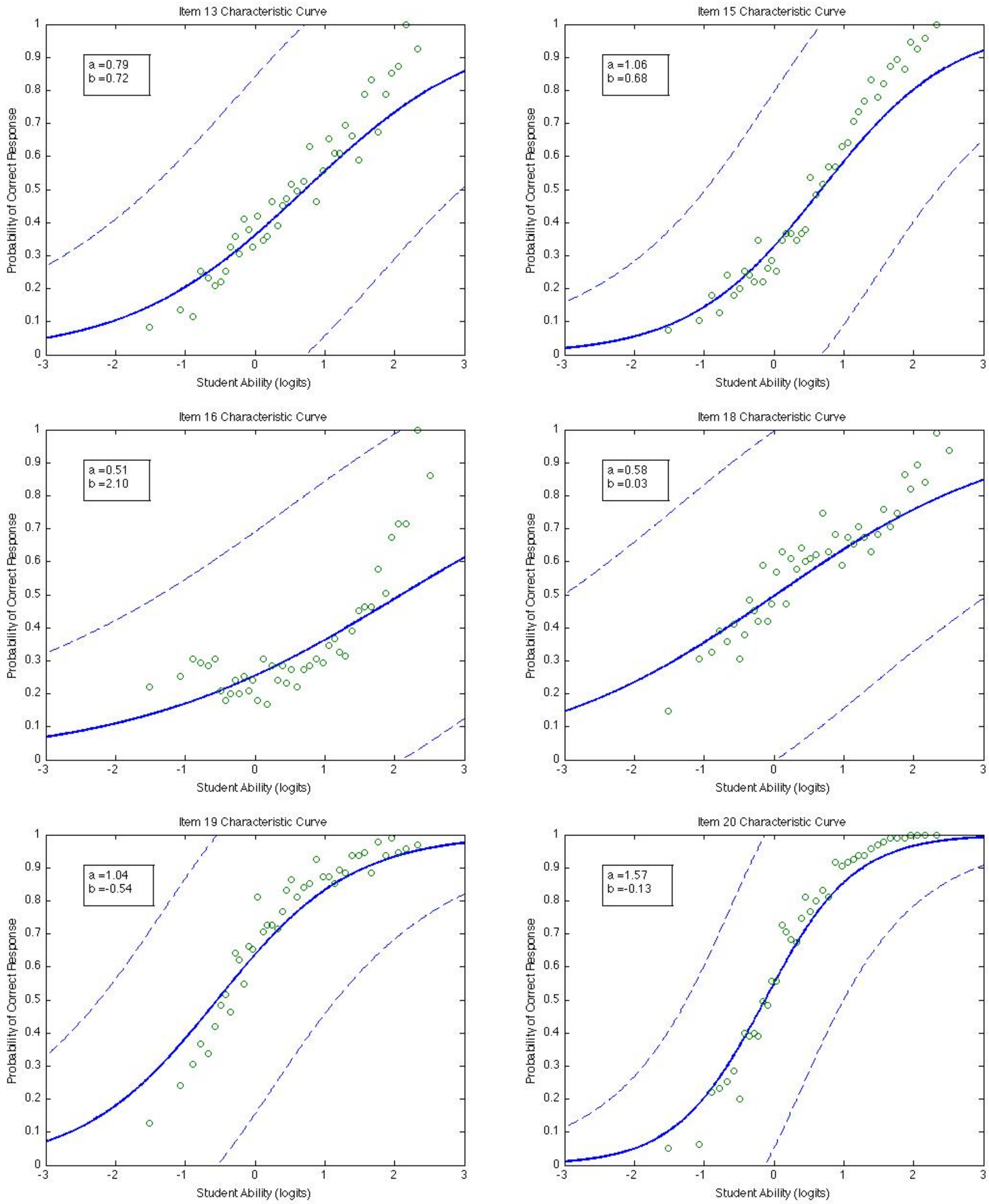

Item 6 Characteristic Curve
a =0.70
b =-0.18
Probability of Correct Response
Student Ability (logits)


Item 8 Characteristic Curve
a =1.79
b =0.79
Probability of Correct Response
Student Ability (logits)

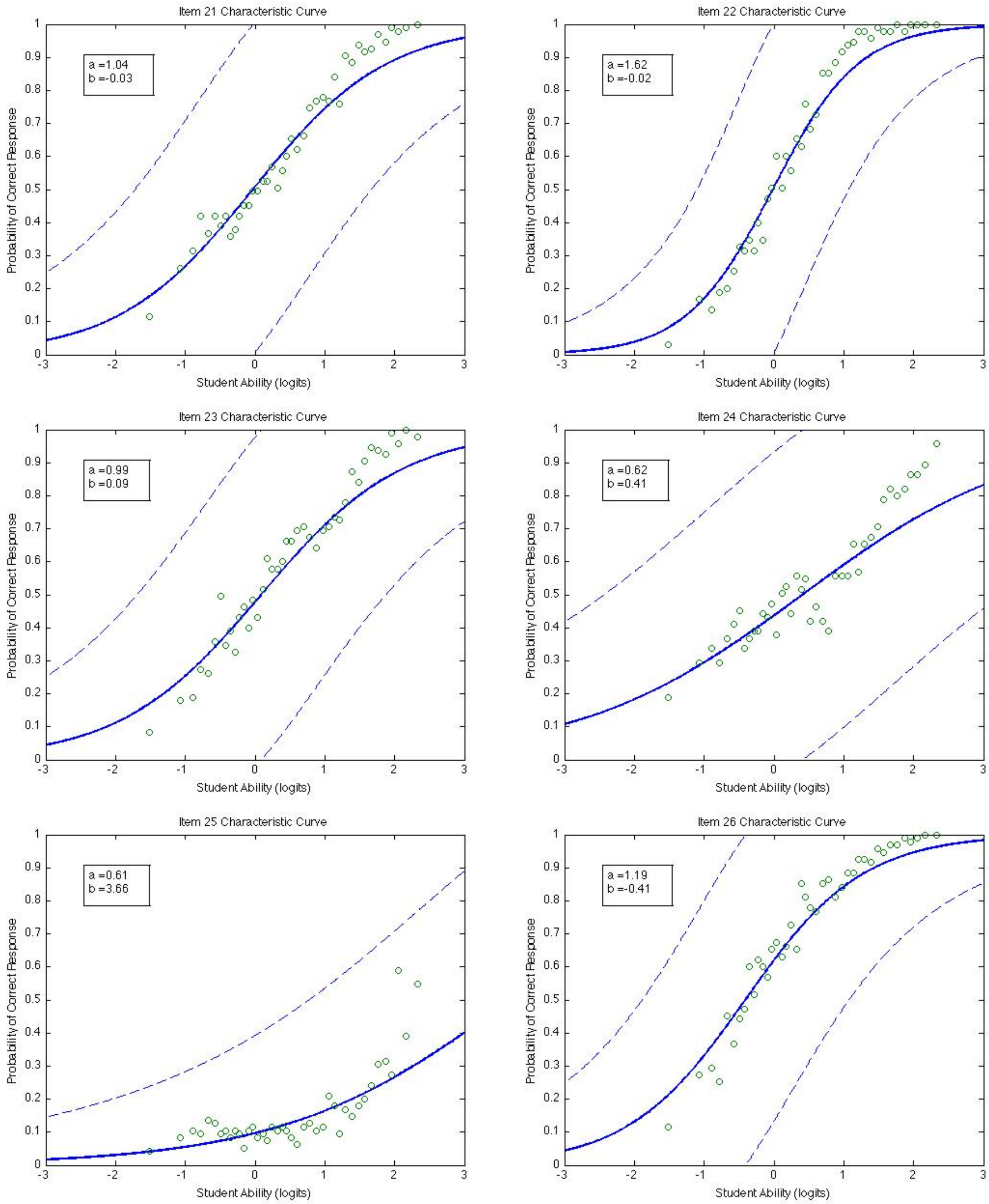

Item 9 Characteristic Curve
a =0.62
b =-1.18
Probability of Correct Response
Student Ability (logits)


Item 11 Characteristic Curve
a =0.69
b =-0.63
Probability of Correct Response
Student Ability (logits)


Item 12 Characteristic Curve
a =0.98
b =-0.26
Probability of Correct Response
Student Ability (logits)



Item 13 Characteristic Curve
a =0.79
b =0.72
Item 15 Characteristic Curve
a =1.06
b =0.68
Item 16 Characteristic Curve
a =0.51
b =2.10
Item 18 Characteristic Curve
a =0.58
b =0.03
Item 19 Characteristic Curve
a =1.04
b =-0.54
Item 20 Characteristic Curve
a =1.57
b =-0.13
Probability of Correct Response
Student Ability (logits)


Item 21 Characteristic Curve
a =1.04
b =-0.03
Item 22 Characteristic Curve
a =1.62
b =-0.02
Item 23 Characteristic Curve
a =0.99
b =0.09
Item 24 Characteristic Curve
a =0.62
b =0.41
Item 25 Characteristic Curve
a =0.61
b =3.66
Item 26 Characteristic Curve
a =1.19
b =-0.41
Probability of Correct Response
Student Ability (logits)

Calibrating the NGCI

The IRT parameters represent a robust “fine-tuning” of the NGCI instrument calibration process, broadening the applicability of the NGCI as a standardized measurement tool. Just as a meter stick does not depend on the object whose length it measures, the parameter invariance of IRT means that the item parameters in Table 24 can be used to measure understanding of Newtonian gravity for any Astro 101 and Phys 101 student. To get a better idea of how IRT parameter estimates standardize the items for both Astro 101 and Phys 101 populations, Figure 8 and Figure 9 compare IRT parameters to post-instruction astronomy and physics CTT parameters. While CTT and IRT item parameters are not on the same scale, items in Figure 8 are ordered from least difficult to most difficult, and items in Figure 9 are ordered from least discriminating to most discriminating, providing some insight into how the item functionality discussed in Chapters 4 and 5 compares to the IRT results.

Inspecting Figure 8 shows that, in general, the ordering of IRT item difficulty parameters map to a combination of those from each group’s CTT difficulty parameters. For example, the two items with the lowest CTT difficulty for the post-instruction astronomy sample were Items 9 and 11, while for the post-instruction physics sample they were Items 5 and 26. The IRT difficulty estimates show that, overall, these four items are some of the easiest items on the NGCI. Also, items that require proportional reasoning in the Force Law domain (Items 4, 10, 16, and 25) show the highest CTT difficulty values for both groups and the highest overall IRT difficulty parameters (recall Item 10 was eliminated from the IRT analysis to maintain local independence).

Furthermore, recall that IRT difficulty and student ability are measured on the same scale of logits. Table 24 shows that 21 out of the 22 items used for the analysis have *b* values ranging from -1.18 to 2.10, which are more or less equally spaced over a similar range as the ability levels of the students, which range from -2.21 to 2.56 logits, indicating that the NGCI items are well-matched to the target population, maximizing their potential (Planinic, Ivanjek, & Susac 2010).

With an IRT *b* value of 3.66, Item 25 is over a full logit higher than all other items' difficulty values and is the only item whose difficulty is outside the ability range of all the students in the sample. Even students with the highest ability (2.56 logits) have only a 34% probability of answering Item 25 correctly, which may explain why this item has both infit and outfit values above 1.00. This is perhaps not surprising, since the CTT values also indicated that Item 25 is extremely difficult. Item 25 is even difficult for some experts: only 20 out of the 24 (83%) advanced physics students and experts who took the NGCI answered this item correctly. However, as was argued in Chapter 4, Item 25 represents an important probe of students' ability to reason with mass and distance simultaneously. While it is important for the difficulty of most items to be targeted to the ability of the sample population, Item 25 is so advanced that it probes Newtonian gravity ability beyond the sample population.

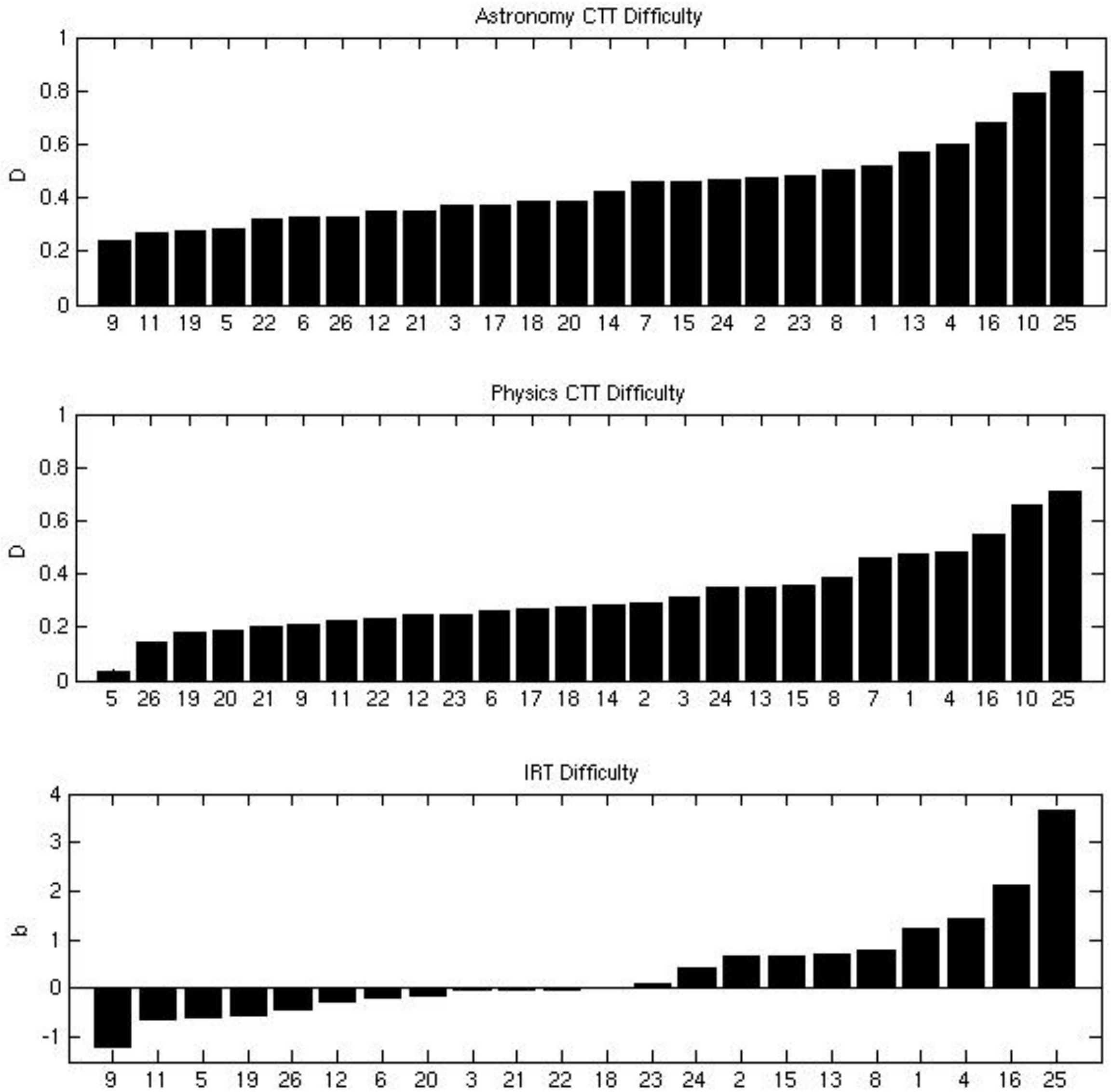


Figure 8: Item difficulty values ordered from easiest to hardest for CTT post-instruction astronomy (top), CTT post-instruction physics (middle), and the four-group 2PL IRT model (bottom).

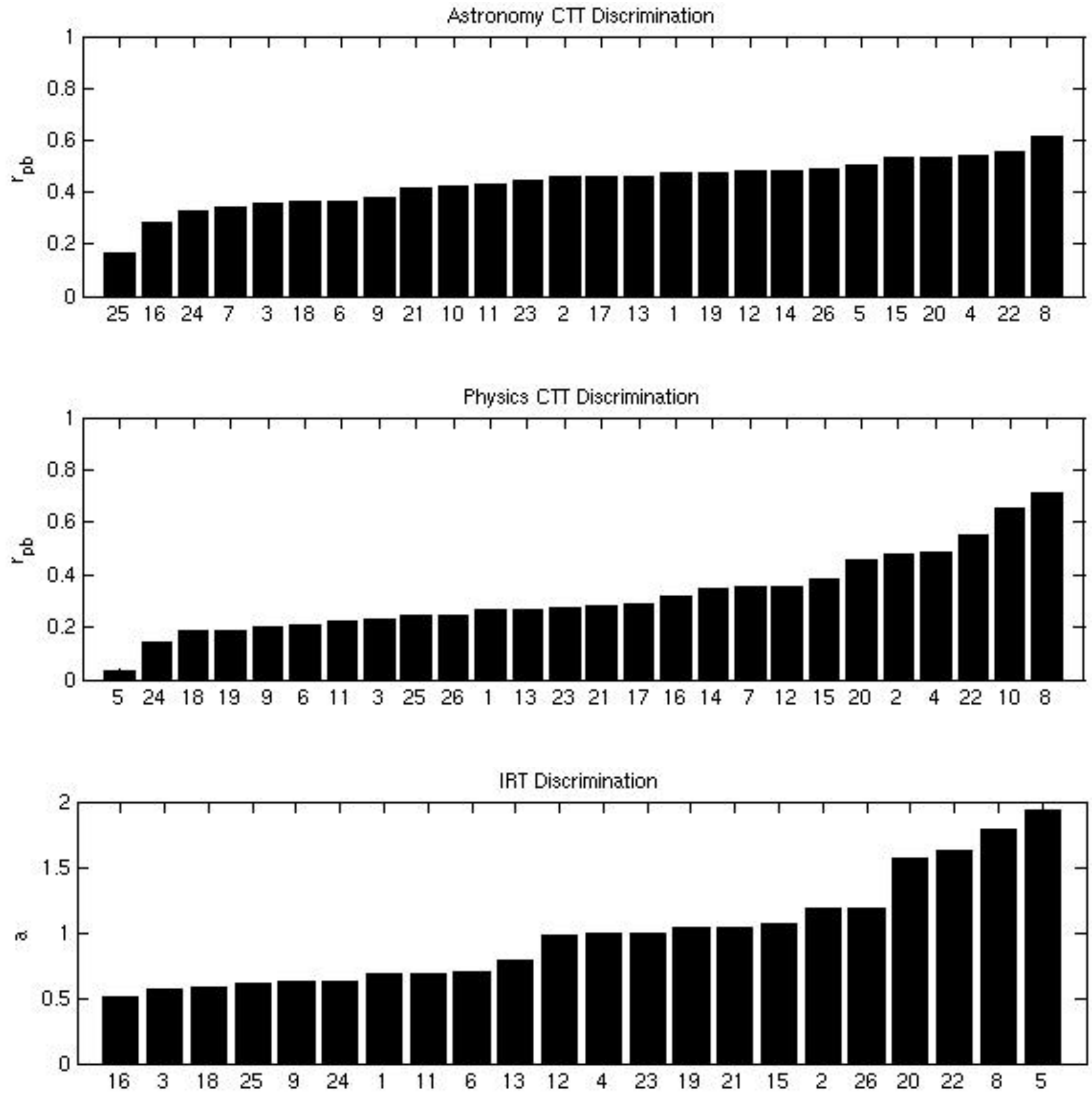


Figure 9: Item discrimination values ordered from least to most discrimination for CTT post-instruction astronomy (top), CTT post-instruction physics (middle), and the four-group 2PL IRT model (bottom).

Table 24 shows that Items 16, 3, 18, and 25 are the poorest IRT discriminators of student ability (they have rather flat ICCs) and Items 22, 8, and 5 are the best discriminators of student ability (they have steeper ICCs). Additionally, Figure 9 shows that, while this trend is somewhat evident in the ranking of CTT discrimination parameters, the ordering of most items' IRT discrimination parameters does not necessarily map to that of each group's CTT discrimination values. Item 5 is perhaps the most glaring example of this. Item 5 shows a somewhat high CTT discrimination value for the post-instruction astronomy group and it is the poorest discriminator of post-instruction physics student performance, but it shows the highest IRT discrimination parameter, *a*. Why would Item 5 show higher discrimination in IRT than in either group's CTT values? Recall from Figure 4 that Item 5 functions most differently in terms of its physics compared to astronomy CTT difficulty, with 72% and 97% of astronomy and physics students answering correctly, respectively. Item 5, therefore, polarizes physics and astronomy students, explaining why it is a good IRT discriminator. Plotting the Item 5 ICC with overlaid student data separated by group (Figure 10) shows the stark contrast between physics and astronomy students, with the post-instruction physics data points populating only the upper tail of the S-curve. In fact, it is only the astronomy student data that maps the entirety of the probability function. There is very little deviation from this general "rule of thumb" that physics students mostly get the item correct and astronomy students show a range of probabilities for answering correctly, perhaps explaining why the outfit value for Item 5 is so low. Furthermore, recall that Item 26 is conceptually similar to Item 5 and shows the second lowest CTT difficulty value ($D = 0.14$) for post-instruction physics students. Its IRT discrimination value is also ordered higher than either the astronomy or physics CTT values, with the ICC showing a similar bifurcation between physics and astronomy students, although not as pronounced as that for Item 5 (Figure 10).

Recall from Chapter 5 that the next five most polarizing items that separate astronomy students compared to physics students fall within the Threshold domain: Items 2, 13, 20, 21, and 23. While the ICCs with overlaid group data for these items (Figure 11) do not show as stark a contrast as those for Items 5 and 26, one can still see a separation between the astronomy and physics data points, with the physics data shifted slightly to higher ability and probability of answering each question correctly. This supports the

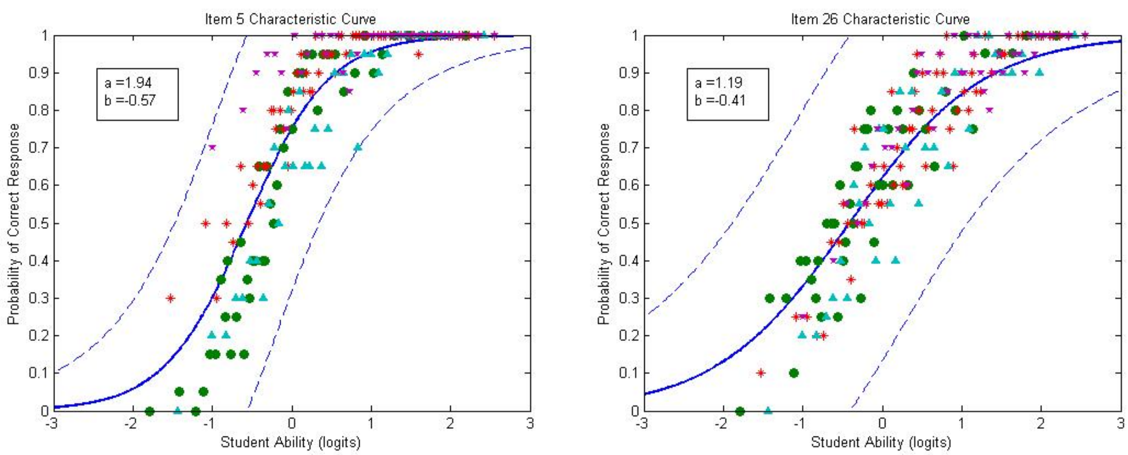


Figure 10: Group data overlaid on the IRT Item Characteristic Curves show that the high IRT discrimination indices for Items 5 and 26 indicate a bifurcation between astronomy and physics students. Green circles represent pre-instruction astronomy students, red asterisks represent pre-instruction physics students, turquoise triangles represent post-instruction astronomy students, and purple stars represent post-instruction physics students. Each data point represents approximately 20 students.

claim that physics students excel most above astronomy students in reasoning with the universality of gravity in limiting cases of low masses and/or large distances.

Finally, while the analysis of IRT parameters for the items shows that the 2PL IRT analysis of the 22-item subset is sensitive to differences *between* physics and astronomy students, it is also important that IRT item parameters be able to measure differences *within* the populations. Group data overlaid on the ICCs for all other items (APPENDIX V) show a mixed distribution of students in both groups with a range of abilities and probabilities for answering items correctly, rather than clear bifurcation between physics and astronomy students. This supports some of the findings from Chapter 5 that physics students approach many NGCI items similarly to astronomy students and, therefore, tend to exhibit the same types of reasoning difficulties as astronomy students (i.e. some physics students struggle with the Boundary Model, the Mixing of Forces model, and misapplications of the scientific model, though to a lesser extent than astronomy students). Having an assortment of items that can measure differences both between and within groups builds on the conclusions of Chapter 5: (1) the NGCI is reliable and valid for *both* astronomy and physics students, and, furthermore, (2) a 22-item subset of the NGCI (excluding Items 7, 10, 14, and 17) can *measure* differences in student understanding along a continuous construct, just like any other physical measurement device. The IRT item parameter estimates represent a more holistic, "big picture," population-independent calibration of the NGCI. This calibration makes the NGCI widely applicable and "ready-to-use" for measuring students' abilities related to Newtonian gravity in any other representative sample population.

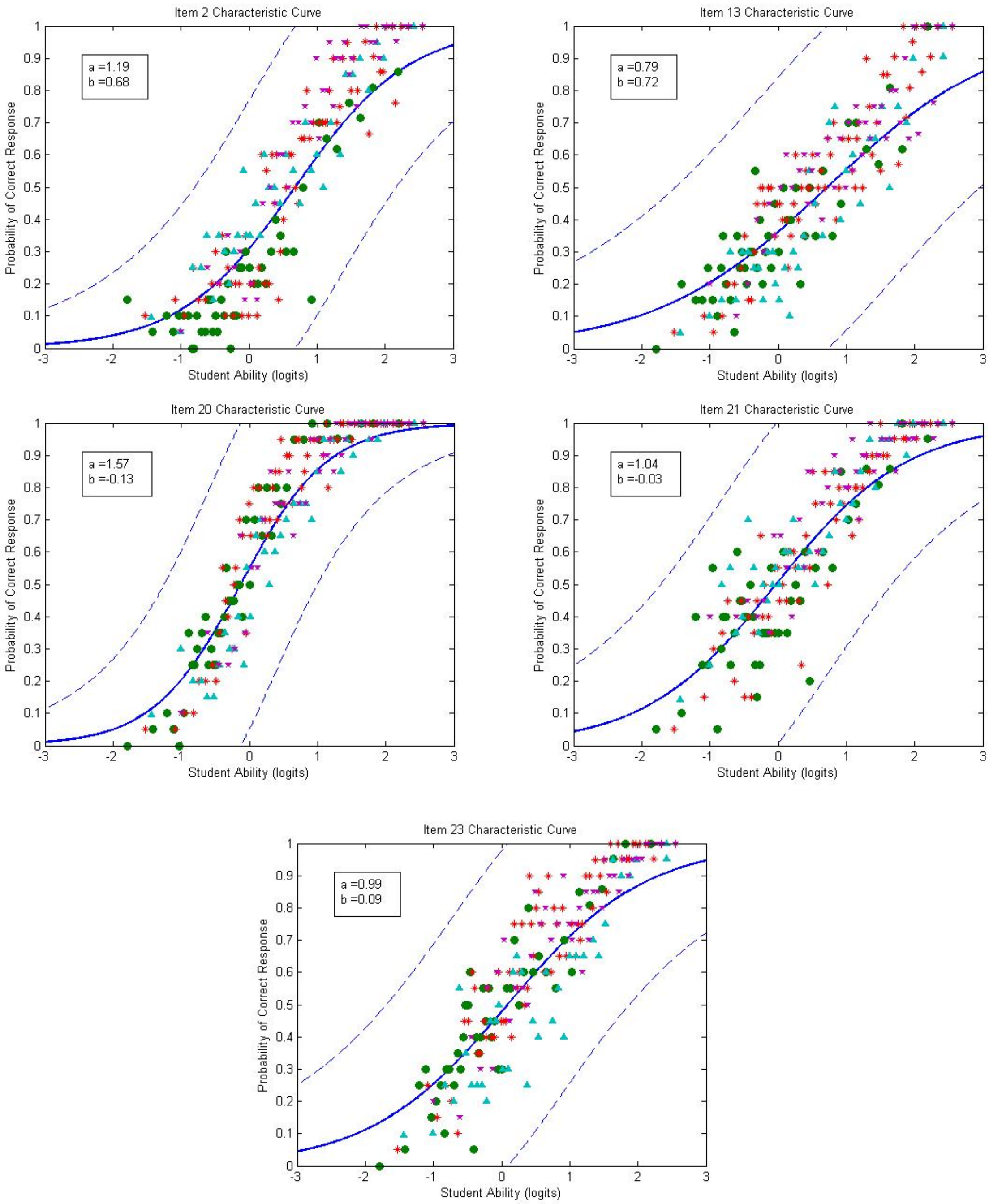


Figure 11: Group data overlaid on the IRT Item Characteristic Curves for items in the Threshold domain indicate a distinct shift of physics student data toward higher abilities and probability of answering correctly. Green circles represent pre-instruction astronomy students, red asterisks represent pre-instruction physics students, turquoise triangles represent post-instruction astronomy students, and purple stars represent post-instruction physics students. Each data point represents approximately 20 students.

Measuring Student Ability

The 2PL IRT student ability estimates represent students' "true" abilities along the linear scale of the Newtonian gravity construct. What does this really mean? First, IRT "true" abilities are different than those of CTT, where a student's true score is defined as the expected score a student would get if the student took the same (or a parallel) test over many administrations (recall that this is unknowable). The IRT ability is completely independent of the instrument. For example, a student with an IRT ability estimate of 0.50 logits should display an ability of 0.50 logits on *any* reliable and valid tool for measuring student understanding of Newtonian gravity, even if that tool is much easier or much harder than the NGCI (Hambleton & Jones 1993). Furthermore, measuring student ability on a linear scale of logits allows direct comparisons between students to be made. One *cannot* say, for example, that a student with an NGCI score of 80% possesses twice the understanding of another student who has a score of 40%, but with IRT one *can* say that a student with an ability of 0.80 logits possesses twice the understanding of a student with an ability of 0.40 logits (Planinic, Ivanjek, & Susac 2010, Wallace & Bailey 2010). Furthermore, 2PL IRT ability estimates account for item parameters, giving more weight to correct responses to higher discriminating items, so even students who have the same raw score may not have the same ability (Harris 1989, Wallace & Bailey 2010), providing a "fine-tuning" of the interpretations one can make about students. This section leverages these important facets of the 2PL IRT ability estimates to explore the differences between physics and astronomy students' Newtonian gravity ability both pre- and post-instruction, in addition to students' gains in ability as the result of the physics compared to astronomy curriculum.

First, histograms of physics and astronomy students' IRT ability estimates (Figure 12 and Figure 13) clearly show that physics students as a whole have higher Newtonian gravity ability than astronomy students both pre-instruction (Figure 12) and post-instruction (Figure 13). Using each student's IRT ability estimate and accompanying weighting factor, w_p, computed from the standard error returned by BILOG-MG ($w_p = 1/SE_p$), each group's average ability, $\hat{\theta}$, could be calculated:

$$\hat{\theta} = \frac{\sum_{p=1}^{n} w_p \theta_p}{\sum_{p=1}^{n} w_p}$$

along with an uncertainty, $\hat{\sigma}$ (Taylor 1982 via Wallace 2011):

$$\hat{\sigma} = \left(\sum_{p=1}^{n} w_p\right)^{-1/2}$$

The average ability of the 925 Astro 101 pre-instruction students was -0.05 ± 0.02 logits, which was *0.64 ± 0.28 logits lower* than the 1,392 Phys 101 pre-instruction students' average ability (0.59 ± 0.02 logits). The average ability of the 743 Astro 101 post-instruction students was 0.43 ± 0.02, which was *0.56 ± 0.28 logits lower* than the 929 Phys 101 post-instruction students' average ability (0.99 ± 0.02 logits). Error estimates for differences in ability were computed by adding those of the average ability estimates in quadrature (Taylor 1982 via Wallace 2011). One-tailed independent samples *t*-tests (equal variances assumed) show that these observed differences in ability were significant, with physics students' abilities significantly higher than astronomy students' abilities both pre-instruction ($t(2315)=-17.83$, $p<0.001$) and post-instruction ($t(1670)=-13.00$, $p<0.001$).

Taken together, the CTT analysis, distribution of response patterns, and the 2PL IRT analysis of students' abilities offer a compelling argument that, in general, physics and astronomy students start and finish their courses with very different understandings of Newtonian gravity. But, what are the effects of instruction on student *learning* of Newtonian gravity? One will note that the average post-instruction abilities of both populations are higher than their average pre-instruction abilities. Are there differential effects of the physics compared to the astronomy courses in the sample? Hake's (1998) gain is typically used to assess learning, but this method has some known shortcomings. First, it depends on students' scores, which provide only ordinal measures of student understanding. Second, it is biased toward students with high pre-instruction scores. And third, the formula fails to provide meaningful interpretations when post-instruction scores are lower than pre-instruction scores (Wallace & Bailey 2010). With IRT abilities measured on a linear scale, however, gains in IRT are much more straightforward. IRT gains are simply the difference between students' average pre- and post-instruction abilities (Hambleton & Jones 1993), again with error estimates added in quadrature (Taylor 1982 via Wallace 2011). For example, a student with a pre- and post-instruction ability of -1.20 and -0.20 logits, respectively, gained the same amount of Newtonian gravity ability as a student who had, say, a pre- and post-instruction ability of -0.10 and 0.90 logits. Both students gained 1.0 logits of ability. These kinds of statements are not possible with traditional total scores, again showing the robust power of IRT. Table 26 shows the averaged IRT abilities and gains for each Astro 101 and Phys 101 class in the sample, providing insight into differences in instruction and curriculum.

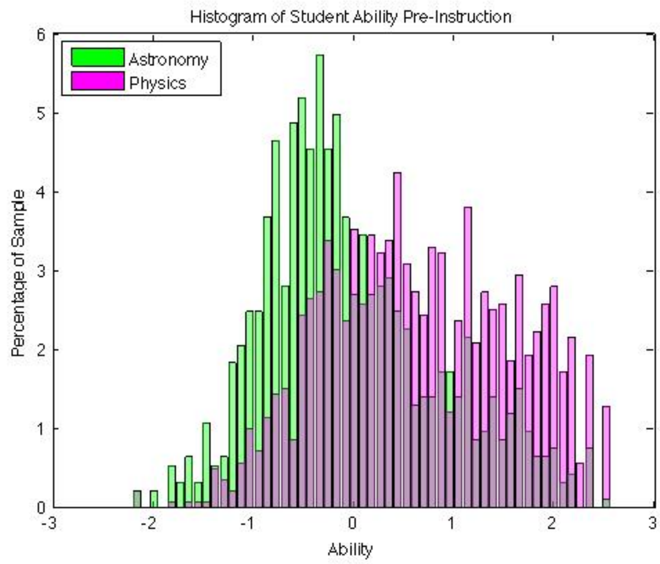


Figure 12: Histogram of physics and astronomy students' ability pre-instruction.

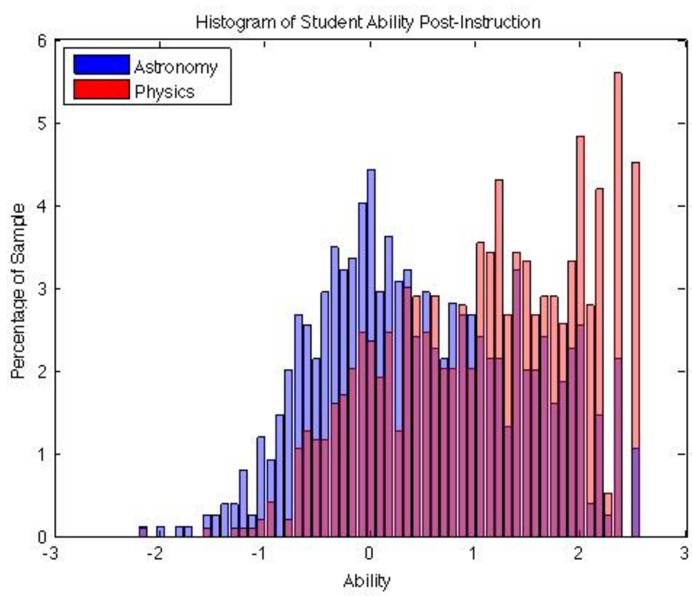


Figure 13: Histogram of physics and astronomy students' ability post-instruction.

Table 26: Class averaged pre-instruction ability, post-instruction ability, and IRT gains.

Institution	Year	Mean Pre-Instruction IRT Ability	Mean Post-Instruction IRT Ability	IRT Gain
Astronomy				
Montana State University	2012	-0.04 ± 0.04	0.18 ± 0.04	0.22 ± 0.06
University of Arizona	2012	-0.18 ± 0.04	0.41 ± 0.04	0.59 ± 0.06
Westchester CC	2012	-0.33 ± 0.08	0.39 ± 0.04	0.72 ± 0.09
Youngstown State Univ.	2012	-0.34 ± 0.06	0.38 ± 0.09	0.72 ± 0.11
Truckee Meadows CC	2012	-0.05 ± 0.16	1.06 ± 0.24	1.11 ± 0.29
UC Davis Solar System	2012	0.61 ± 0.10	0.71 ± 0.05	0.10 ± 0.11
UC Davis Stars and Galaxies	2012	0.48 ± 0.06	1.11 ± 0.07	0.63 ± 0.09
Physics				
Montana State University	2012	n/a	0.78 ± 0.08	n/a
	2013	0.34 ± 0.04	0.75 ± 0.05	0.41 ± 0.06
College of Dupage	2012	1.04 ± 0.10	1.64 ± 0.13	0.60 ± 0.16
Northern Arizona University	2012	0.23 ± 0.08	0.83 ± 0.09	0.60 ± 0.12
	2013	-0.01 ± 0.05	0.28 ± 0.05	0.29 ± 0.07
University of Maine	2012	0.73 ± 0.06	n/a	n/a
	2013	0.59 ± 0.04	1.27 ± 0.06	0.68 ± 0.07
UC Santa Barbara	2013	1.17 ± 0.04	1.71 ± 0.05	0.54 ± 0.06
Buffalo State University	2013	0.30 ± 0.12	1.13 ± 0.15	0.83 ± 0.19
Snow College	2013	0.40 ± 0.15	1.13 ± 0.17	0.73 ± 0.22
Kilgore College	2013	0.00 ± 0.14	n/a	n/a

Table 26 shows that, while these IRT gains echo the Hake gains presented in Table 5 and Table 17 (i.e. lower IRT gains correspond to lower Hake gains, and higher IRT gains correspond to higher Hake gains), there is not necessarily a one-to-one correspondence. The wide range of IRT gains show differences between courses, but, with a range from 0.10 ± 0.11 to 1.11 ± 0.29 for the Astro 101 courses and from 0.29 ± 0.07 to 0.83 ± 0.19 for the Phys 101 courses, it is difficult to determine if there is a differential impact of Astro 101 compared to Phys 101 curriculum (i.e. the different topical contexts in which Newtonian gravity is taught) on student learning of Newtonian gravity. Furthermore, besides being in an astronomy or a physics class, there may be a variety of other factors that influence the observed student IRT gains, including individual students' characteristics (for example, such as interest in the subject, motivation to complete the NGCI to the best of their ability, gender, experience in previous physics or astronomy courses taken, academic major, mathematical preparation, socioeconomic status, parents' education level, and ethnicity, among many other aspects of the student) and classroom dynamics (such as class size, time allocated for administering the NGCI, whether the course was taken during the Fall or Spring semester, the physical space of the classroom, instructor's teaching experience, institution type, time spent teaching about Newtonian gravity, and prevalence of interactive engagement). Similar factors have been shown to contribute to variances in student scores on the Force Concept Inventory and the Light and Spectroscopy Concept Inventory, with the degree of interactive engagement practices being the most prevalent indicator of student performance (Meltzer & Manivannan 1996, Hake 1998, Prather et al 2009, and Rudolph et al 2010). To what extent do these factors influence the NGCI results on the differences between astronomy and physics student ability estimates? In controlling for some of these factors, is it possible that a differential effect of the Astro 101 compared to the Phys 101 curriculum would emerge? To help answer these questions, the following section uses multiple linear regression models to predict students' IRT gains while controlling for some student characteristics and classroom dynamics.

Regression Analysis of IRT Gains

Multiple linear regression analysis is a versatile and powerful statistical technique that allows researchers to examine the strength, direction, and statistical significance of the association between a dependent variable and a predictor variable while controlling for other predictor variables that may confound this association (Urdan 2010). The dependent variable, $\hat{Y}$, is predicted to be a linear combination of k variables, X, with weighted contributions, B, plus a constant (Brace, Kemp, & Snelgar 2013):

$$\hat{Y} = constant + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

The constant B_i is the correlation coefficient for the dependent variable and independent variable X_i times the ratio of the standard deviations, thereby transforming the correlation coefficient into the scales of measurement used for the two variables (Urdan 2010). Scale predictor variables must be on an interval scale and nominal predictor variables must be dichotomously dummy coded, such that a predictor variable's B value indicates how much *Y* will change if the predictor changes by one unit. Because each predictor variable may be on a different scale, these B values are typically "standardized" into beta values, which allow one to compare the relative strengths of different predictor variables (Urdan 2010).

In this section, multiple linear regression models are used to explore the effects of Astro 101 courses compared to Phys 101 courses on student learning of Newtonian gravity. IRT gain (i.e. post-instruction ability minus pre-instruction ability) was chosen as the dependent variable, *Y*, with a set of eight predictor variables that control for student demographics and prior experience as well as classroom dynamics and institutional environment. The coding of these variables is discussed below, followed by an interpretation of the results of the multiple regression analyses.

Coding of Variables

The eight predictor variables used for the regression analysis of student IRT gains are shown in Table 27. The primary, or "main effect," predictor of interest is the effect that physics curriculum compared to astronomy curriculum (i.e. the topical context in which Newtonian gravity is taught) has on student learning. Student characteristics were modeled with four predictor variables (Pre-Instruction Ability, Major, Gender, and Previous Courses Taken), and classroom environment was modeled with three independent variables (Years Teaching, Institution Type, and Interactivity Score). The way these variables were determined is discussed below.

Table 27: Coding of variables for regression analysis.

Independent Variable	**Type**	**Dummy Code**
Primary Variables of Interest		
Curriculum	Nominal	Physics = 0 Astronomy = 1
Student Characteristics		
Pre-Instruction Ability	Scale	
Major	Nominal	Non-science = 0 Science = 1
Gender	Nominal	Male = 0 Female = 1
Previous Courses Taken	Nominal	No previous courses = 0 One or more = 1
Classroom Environment		
Years teaching	Scale	
Institution Type	Nominal	2-year = 0 4-year or Research = 1
Interactivity Score	Scale	

In order to track IRT gains in the context of specific student characteristics, only a subset of students who had matched pre-instruction and post-instruction ability estimates were used. This subset consisted of 490 Astro 101 students and 653 Phys 101 students. The student characteristics for this subset were collected from pre-instruction responses to NGCI Questions 29-31 and are shown in Table 28 below. They appear sufficiently similar to the whole group percentages, and, consistent with the observed trends in the whole data set, one-tailed independent samples *t*-tests show that the matched physics subgroup has a significantly higher mean Newtonian gravity ability pre-instruction ($t(1094.050) = 11.350$, $p < 0.00025$, equal variances not assumed) and post-instruction ($t(1141) = 10.625$, $p < 0.00025$, equal variances assumed) than the matched astronomy subgroup. It is assumed this subset is representative of the students in the whole group sample.

Table 28: Matched subset demographic data.

		Astronomy	Physics
Major	Business	22.9%	0.8%
	Education	10.2%	2.3%
	Humanities, Social Sciences, Arts	30.1%	3.5%
	Science, Engineering, Architecture	20.4%	85.1%
	Other	15.5%	6.7%
Previous Courses	0	59.4%	35.5%
	1	29.0%	44.1%
	2 or more	10.2%	17.3%
Gender	Male	45.7%	59.5%
	Female	54.3%	40.5%

In order to use the student demographic data from Table 28 for the regression analysis, some of the demographic categories needed to be collapsed. While academic major initially included five categories, Table 28 shows that the primary difference in major for the physics and astronomy matched subgroups was for the "Science, Engineering, and Architecture" (or succinctly "Science" for the purposes of the regression analysis) category, so the Major variable in Table 27 represents only "Non-science" compared to "Science" majors. The demographic category of Previous Courses Taken was collapsed into "No previous courses" and "One or more previous courses" for similar reasons.

While student characteristics were tabulated from the Questions 29-31 on the NGCI, information about instructional environment was collected with a separate NGCI Instructor Survey, shown in APPENDIX W. This survey was adapted from that of Prather et al (2009), which was used to assess interactive engagement practices of professors who had administered the Light and Spectroscopy Concept Inventory in Astro 101 courses. The NGCI Instructor Survey asked for the institution type (2-year, 4-year, or

Research University) of the pilot site, the number of years the professor had been teaching, and information about the degree of interactivity in the classroom environment. The analysis of these surveys is discussed below and the results for the courses that have matched pre-instruction and post-instruction student data are shown in Table 29.

As in Prather et al (2009), the Interactivity Score can be thought of as the fraction of a course's time that was spent on interactive engagement compared to the total time available in the semester. However, unlike Prather et al (2009), the total available time for instruction was represented as the sum total of time spent in lecture *and* recitation and laboratory. In-class time spent on exams was subtracted. In this sense, the Interactivity Score is modified to include *any* "non-traditional" instruction, i.e. instruction other than lecture and recipe or confirmation "cookbook" laboratories. This accounts for the reality that Phys 101 courses generally require more time investment from students, thereby possibly leading to greater conceptual understanding.

Table 29: Results of the NGCI Instructor Survey for classes with matched pre-instruction and post-instruction student data.

Institution	Type	Years Teaching	Interactivity Score
Astronomy			
Montana State University	4-year	8	20.6
University of Arizona	4-year	20	38.3
Westchester Community College	2-year	10	37.5
Youngstown State University	4-year	8	34.2
UC Davis Solar System	Research	6	34.5
UC Davis Stars and Galaxies	Research	6	37.3
Physics			
Montana State University	4-year	24	39.3
Northern Arizona University	Research	3	25.7
University of Maine	Research	20	29.0
UC Santa Barbara	Research	33	68.4
Buffalo State University	4-year	22	62.2
Snow College	2-year	21	17.7

In calculating the time spent on interactive engagement, collaborative Think-Pair-Share questions were assumed to take three minutes each, and non-collaborative, individual questions were assumed to take one minute each. However, the time spent on collaborative learning activities was calculated differently for physics and astronomy classes. The *Tutorials in Introductory Physics* were designed to be implemented during an hour's recitation time, whereas those in the *Astronomy Lecture-Tutorials* were designed for approximately fifteen-minute activities during lecture time. Therefore, if a Phys 101 Instructor Survey reported using, for example, five collaborative learning activities throughout the semester, 300 minutes were counted (5 activities x 60 minutes per activity) toward the interactive time estimate, whereas if an Astro 101 Instructor Survey reported that same number of activities, only 75 minutes were counted (5 activities x 15 minutes per activity) toward the interactive time estimate.

Special care was taken on a case-by-case basis to modify the Interactivity Score calculation beyond the above considerations. For example, one instructor noted that he used a "flipped" classroom style, in which students viewed a podcast of the lecture before coming to class and worked through homework during class. In this case, an extra class period's time of "instruction" time was added to the actual class meeting time each week. The Interactivity Score for this class is still quite high, echoing the fact that these students indeed received more non-traditional interactive engagement than students in non-flipped environments. In another case, the professor provided time estimates for Think-Pair-Share and individual questions that differed from those above, so an average was taken. Furthermore, there was variation in responses to Question 8 on the Instructor Survey, which asked about "other" non-traditional things that promote learning, requiring some degree of subjective interpretation. The Interactivity Score should, therefore, be viewed as a "first order" estimate. However, the wide range of Interactivity Scores in Table 29 (17.7 to 68.4) indicates that the scoring method that was implemented is sensitive to variations in classroom activity (Prather et al 2009). Furthermore, given that the values in Table 29 were carefully coded and seem qualitatively reasonable with the given information, it is assumed that they are representative of the true classroom environments.

Finally, it is important to note that the Interactivity Scores may not be representative of *all* physics and astronomy courses. For example, one will note that the Interactivity Scores shown in Table 29 tend to cluster near 35%, meaning that 35% of the available instruction time is interactive. This is much higher than purely traditional courses, so it is suspected that this cluster of scores shows that the instructors who volunteered to participate in piloting the NGCI in their classrooms tended to have an interest in improving their teaching and were more likely to implement more non-traditional methods. It should also be noted that the average Interactivity Score for the astronomy classes in the sample was 33.7%, while it was 40.4% for the physics classes. Since the physics courses in the sample had more available instruction time with labs that could be devoted entirely to interactivity, they show a systematically higher Interactivity Score. Again, this is likely not representative of all physics courses, in which purely traditional lecture with "cookbook" labs would still produce an Interactivity Score of zero. These caveats prove important in interpreting the results of the regression analysis models below.

Model Building Approach

In order to use multiple regression, the predictor variables in Table 27 should represent sufficiently different factors so that they each independently contribute to predicting the dependent variable. If the variables are highly correlated with each other, the so-called "multicollinearity" makes interpretations of the regression model uncertain (Brace, Kemp, & Snelgar 2013). Therefore, before implementing regression analyses, the Pearson correlation coefficient was calculated for each pair of predictor variables using the IBM SPSS statistics software package. All pairs showed low or medium correlations, except for Years Teaching and Interactivity Score (0.813) and Curriculum and Major (-0.650). However, Field (2005) and O'Brien (2007) suggest that this high correlation is only a problem if the regression models' multicollinearity diagnostics show Tolerance values below 0.2, or, equivalently, variance inflation factors (VIFs) greater than 5. None of the highly correlated variables were outside of this range for any of the regression models implemented so all variables were retained for the analysis. Furthermore, for the purposes of this study, Curriculum and Major are of primary interest, so it was important to include both variables in the analysis.

To extract as much information as possible about the influence of these eight predictor variable on students' post-instruction Newtonian gravity ability, a multiple linear regression model building approach was adopted in which blocks of control variables were added sequentially. The model-building steps followed the general structure:

Model 1: Y = Curriculum + Pre-Instruction Ability

Model 2: Y = Model 1 + student demographic variables

Model 3: Y = Model 2 + classroom environment variables

The IBM SPSS statistical software package was used to implement each regression model. The SPSS analysis used a least squares approach to calculate the coefficients that maximize the correlation between the observed and model-predicted post-instruction student ability (Brace, Kemp, & Snelgar 2013). The multiple correlation coefficient, R, is a measure of this correlation, and R^2 indicates the proportion of the variance in the dependent variable (i.e. students' IRT gains) that can be explained by the model. Furthermore, an Adjusted R^2 value accounts for the number of predictor variables and the number of students in the sample, providing the most useful measure of model success (Brace, Kemp, & Snelgar 2013).

Results

The results of the three regression models are shown in Table 30. One will note that each model shows significant R^2 and Adjusted R^2 values ($p<0.001$), indicating that all the models adequately match the observations. And, the variance explained significantly increases with each model ($p<0.001$), i.e. with each addition of blocks of control variables, indicating that the models become progressively more robust in predicting students' gains in Newtonian gravity ability. What do the results of these models mean?

Table 30: Multiple linear regression coefficients of NGCI IRT Gain on Curriculum.

	Model 1		Model 2		Model 3	
Independent Variable	**B (SE)**	**β**	**B (SE)**	**β**	**B (SE)**	**β**
Constant	0.800** (0.035)		0.670** (0.073)		0.259* (0.125)	
Curriculum (astronomy)	-0.262** (0.050)	-0.143**	-0.137* (0.062)	-0.077*	-0.002 (0.065)	-0.001
Pre-Instruction Ability	-0.482** (0.025)	-0.523**	-0.546** (0.029)	-0.600**	-0.592** (0.028)	-0.650**
Major (science)			0.143* (0.064)	0.080*	0.147* (0.065)	0.082*
Gender (female)			-0.213** (0.050)	-0.121**	-0.184** (0.049)	-0.104**
Previous Courses (one or more)			0.198** (0.051)	0.112**	0.074 (0.051)	0.042
Years Teaching					0.007 (0.004)	0.086
Institution Type (4-yr or research)					-0.146 (0.099)	-0.038
Interactive Engagement Score					0.011** (0.003)	0.198**
R^2	0.222**		0.251**		0.303**	
Adjusted R^2	0.220**		0.247**		0.298**	
R^2 change			0.029**		0.052**	
Df	2		3		3	

* p<0.05, **p<0.001

Model 1 includes the main effect variable of Curriculum controlled for Pre-Instruction Ability. Since the physics students started at significantly higher ability levels, this variable was of primary importance to control. Results show these two variables account for 22.2% of the observed variance in IRT gains. The coefficient for Pre-Instruction Ability (B = -0.482, p<0.001) shows that starting with a higher pre-instruction ability leads to significantly lower gain (more about what this might mean is addressed in the Discussion section below). While Pre-Instruction Ability has the greatest weight for predicting gain in student ability (Beta = -0.523, p<0.001), the Curriculum variable is also significant (Beta = -0.143, p<0.001). The Model 1 Curriculum coefficient of B = -0.262 (p<0.001) implies that, for students of equal Pre-Instruction Ability, taking the physics curriculum leads to an IRT gain that is 0.262 logits *higher* than taking astronomy. Given the drastic differences between these two curricula and the types of students who take them, it is natural to wonder if other variables are also influencing this result.

Model 2 includes an additional block of student characteristics control variables, including Major, Gender, and Previous Courses Taken. By including these additional student characteristic variables, Model 2 shows a significant ($p<0.001$) improvement over Model 1 and significantly explains 24.7% of the observed variance in gains. All four of the student characteristic variables of Pre-Instruction Ability, Major, Gender, and Previous Courses Taken are significant factors in predicting a student's gain in ability. As with Model 1, the variable with the largest effect is Pre-Instruction Ability, with higher Pre-Instruction Ability again predicting lower gain in ability. Model 2 also predicts that: science majors will gain 0.143 logits *more* than non-science majors ($p<0.05$) in understanding of Newtonian gravity, females will gain 0.213 logits *less* than males ($p<0.001$), and students who had taken previous courses will gain 0.198 logits *more* than those who had not ($p<0.001$). The effect of Curriculum is still significant, although less significant than in Model 1 (B = -0.137, $p<0.05$), predicting that the physics curriculum still leads to differentially *higher* IRT gains. However, there are still other factors that may influence this gain, including course structure and classroom environment.

Model 3 controls for the additional variables of classroom dynamics on top of student characteristics, including professor's number of Years Teaching, the Institution Type (2-year or 4-yr/Research), and Interactivity Score (Table 29). Including these variables significantly improves the predictive power of the regression model ($p<0.001$), providing a total of 29.8% of variance in IRT gains explained by the regression model. The significant predictors of student IRT gain in Model 3 ordered from strongest to weakest are: Pre-Instruction Ability (beta = -0.650, $p<0.001$), Interactivity Score (beta = 0.198, $p<0.001$), Gender (beta = -0.105, $p<0.001$), and Major (beta = 0.082, $p<0.05$). Again, higher Pre-Instruction Ability predicts lower gains, being female predicts lower gains, and being a science major predicts higher gains. Also, Interactivity Score predicts higher gains, a finding that once again echoes the strong evidence in the literature that the prevalence of interactive engagement in the classroom is one of the most important factors in leading to greater learning (Hake 1998, and Prather et al 2009). The variables of Years Teaching and Institution Type are non-significant. Additionally, by including the block of classroom environment variables over Model 2, the Previous Courses variable and *the Curriculum variable lost significance*. It appears that the systematically higher physics interactivity scores (discussed in the Coding of Variables section above) were confounding the effects of the Curriculum variable in Models 1 and 2. Therefore, unlike Models 1 and 2, Model 3 implies that, *after controlling for student characteristics and classroom environment, there is no longer a statistically significant differential effect of physics compared to astronomy curriculum on student learning of Newtonian gravity*. To clarify, Model 3 implies that, to improve student learning of Newtonian gravity, instructors should focus their instruction on methods that include interactive engagement, build on student's pre-instruction ability, and account for differences in students' gender and major, rather than the topical context in which Newtonian gravity is taught. A discussion of *why* these variables might be so important is included below.

Interpretation

The results of Model 3 indicate that student IRT gains rely first on students' Pre-Instruction Abilities, second on the prevalence of interactivity in the classroom, third on Gender, and fourth on Major, so it is instructive to investigate *why* these variables are so important to student learning, and, furthermore, why the Curriculum variable is not. First, it should be noted that the significance of the Gender variable merely echoes the well-known "Gender Gap" that pervades much of the sciences. Therefore, while it is disappointing that females generally have lower gains in Newtonian gravity ability than males, this difference does not threaten the validity of the NGCI. And, the significance of the Major variable that shows how science majors have higher gains is perhaps "obvious." Therefore, the Gender and Curriculum variables will not be heavily discussed here.

The high-impact significance of the Pre-Instruction Ability variable is perhaps surprising. It is the most important predictor of student learning of Newtonian gravity, with students of *higher* Pre-Instruction Ability having *lower* gains. One might expect students who start the course with a high ability level to more easily adopt an even more sophisticated understanding of Newtonian gravity. However, the *lower* observed gains might be a manifestation of the logic behind why Hake's gain formula favors students of higher ability: starting with a greater understanding of Newtonian gravity may make *gaining* understanding more difficult, because *there is so only so much more one can understand*, i.e. this may simply be a law of diminishing return.

But, when Previous Courses and Major are controlled, what does it mean for a student to have high Pre-Instruction Ability in the first place? Where does this ability come from if not from prior experience and area of interest? One idea is that students with high Pre-Instruction Ability may naturally have better conceptual understanding and reasoning abilities that lead to a more sophisticated and scientific way of knowing (i.e. more of an understanding that a theory of gravity should be universally applicable and an inherent set of internally consistent ways of reasoning about physical phenomena that lead to a higher probability of answering consistently across all NGCI items). These ways of knowing can improve with instruction, perhaps explaining why the lower ability students who are lacking these skills and ways of thinking are the ones who truly advance (i.e. exhibit higher IRT Gains) as a result of instruction, closing the gap between students of differing Pre-Instruction Ability. Interviews suggest that physics students (who start with a higher average Pre-Instruction Ability) do indeed exhibit more of this scientific way of thinking (see students F12-01 through F12-12 in APPENDIX C and APPENDIX S); however, with so few interviews and no clear reference to this found in the literature, it is still difficult to say for sure. Another idea that could explain the particularly high Pre-Instruction Ability level of physics students is that the student characteristics variables included in the regression analysis may not represent all the relevant variables that could influence a student's incoming ability. Considering the wide variety of experiences that can shape students' mental landscapes within the cognitive constructivist framework of learning (discussed in Chapter 1), this explanation must be at least partially applicable, although the extent of its applicability is unclear.

The second most important predictor variable for student IRT gains is the Interactivity Score of the class. It is particularly interesting that including this variable in the regression model caused the Curriculum variable to lose significance. What does this mean about the different effect of physics compared to astronomy topical context in which Newtonian gravity is taught on student learning of Newtonian gravity? First, one will note that physics classes typically have more laboratory time, and, therefore, more total time for instruction. And, as evidenced by their higher Interactivity Scores, the physics classes in the sample used this time wisely (as discussed in the scoring of the Instructor Survey and shown in Table 29), implementing non-traditional, interactive learning strategies in both lecture and lab. Therefore, Model 3 indicates that this additional time spent on interactive engagement, rather than the topical context of the class itself, is what led to greater student learning of Newtonian gravity.

The non-significance of the Curriculum variable after controlling for seven other predictor variables is perhaps surprising. Recall that Astro 101 courses teach Newtonian gravity in the context of tides, orbital dynamics, and the structure of the solar system, galaxy, and universe at large, while Phys 101 courses tend to focus on gravity’s effects on Earth in context with kinematics, Newton’s Laws, weight forces, and dynamics. Both curricula lead to gains in student understanding of gravity, and, as discussed in Chapter 5, sometimes the different curricula influence students’ approaches to the questions within the four NGCI conceptual domains. However, it is concluded that the different contexts of Newtonian gravity as discussed in physics compared to astronomy in lectures, labs, homework problems, and exams throughout the entire course do not have a significantly different effect on students' *overall gain* in Newtonian gravity ability. Further, it is asserted that engaging learners in these different contexts related to Newtonian gravity while using interactive learning strategies is the important factor that leads to increased understanding of the topic.

Discussion

This chapter leverages the "parameter invariance" of the powerful statistical framework of Item Response Theory. The Two-Parameter Logistic IRT function was applied to the dichotomously scored responses of 3,989 Astro 101 and Phys 101 students to a 22-item subset of the NGCI. This subset of items satisfies both the unidimensionality and local independence assumptions of IRT. The 2PL IRT model is shown to fit the data, meaning that the invariance of item parameters calibrates the NGCI for all Astro 101 and Phys 101 students, greatly extending the results of the Classical Test Theory analyses presented in Chapters 4 and 5. Additionally, the invariance of student abilities estimates allows for direct comparisons between physics and astronomy students to be made, including an analysis of student gains. Class-averaged gains are presented and a regression analysis with matched student data controls for student characteristics and classroom dynamics to investigate the effects of the physics compared to astronomy curriculum on student learning of Newtonian gravity, finding that there is no difference in effect of curriculum, while pre-instruction ability and interactivity in the classroom are the largest predictors of student gain.

CHAPTER SEVEN

CONCLUSIONS

The primary purpose of this dissertation is to develop, use, and calibrate a reliable and valid instrument that can measure introductory college astronomy student understanding of Newtonian gravity. This instrument, the Newtonian Gravity Concept Inventory, or NGCI, is a robust tool that can join the suite of Astro 101 concept inventories to be used by educators to inform their teaching and by researchers to investigate the effects of instructional interventions and differences among students. It is shown that the NGCI has wide applicability, including the capacity to measure introductory algebra-based physics (Phys 101) students' understanding of Newtonian gravity. Because Newtonian gravity is a topic that is taught in both introductory astronomy *and* introductory physics, the NGCI can offer a unique, qualitative and quantitative, perspective on the *differences* between these vastly different student bodies and curricula. This chapter provides a culminating summary of the development and use of the NGCI to illustrate the power of this new instrument in providing valuable information about introductory college astronomy and physics student understanding of Newtonian gravity.

Summary

This research is framed within a strong foundation of research on how people learn, specifically building on the theory of Constructivism. Chapter 1 uses

Constructivism to argue for the importance of characterizing and measuring student understanding of Newtonian gravity in the Astro 101 population. It is asserted that learning is maximized when instructors take an approach to teaching that is informed by their students' preconceptions, misconceptions, and learning difficulties that arise in interaction with the curriculum. The large numbers of non-science major students taking Astro 101 make it an important target group in which to instill conceptual change on a broad scale, and the topic of gravity is a concept for which everyone has some intuitive understanding, making it an excellent topic from which to build consistent physical theories of force and motion. A multiple-choice Concept Inventory provides a fast, reliable, and standardized way of assessing students' naïve ideas and learning difficulties and how these change as the result of instruction. Chapter 2 provides background about the topic of gravity, reviewing the historical progression of how philosophers and scientists reasoned about gravity from Aristotle to the medieval era to Newton. To illustrate the importance and applicability of Newton's Universal Law of Gravitation in describing the motion of objects, its conceptual and mathematical consequences are also developed. A background literature review of research on student understanding of gravity shows that in many ways it mimics the early historical progression. Students' understanding of gravity is closely related to their understanding of Earth's shape, why things fall, and how "force" works. Discussions about this literature were divided into four categories that map to the four conceptual domains used to explore student understanding of gravity throughout this project (Table 3). It is noted that very little prior work has been done on college students' understanding of Newtonian gravity, particularly in the Astro 101 population, yet it is taught in both Astro 101 and Phys 101 courses. This research helps to fill this important gap.

In keeping with the Constructivist tradition, Chapter 3 establishes a foundation of known student reasoning patterns when thinking about gravity. Consulting physics, astronomy, and Astro 101 experts helped to define a conceptual focus for three open-response questionnaires (APPENDICES E-F). Through a constant comparative, phenomenographic analysis of students' written responses and interview data, prominent reasoning patterns emerged. Characterizing these patterns led to the description of three alternative mental models of gravity (the Boundary Model, the Mixing of Forces Model, and the Orbital Indicator Model), as well as misapplications of the scientific model (in regards to mass, distance, and what causes or creates gravity). The most common student responses that manifest these mental models and related student conceptual and reasoning difficulties formed the basis of distractor choices for the development of the multiple-choice concept inventory. Chapter 3 represents the Planning phase of Benson and Clark's (1982) model for instrument development.

Chapter 4 follows the remaining three phases of Benson and Clark's (1982) model – Construction, Quantitative Evaluation, and Validation – to elucidate the creation and evolution of three versions (Version 1 in APPENDIX I, Version 2 in APPENDIX M, and Version 3 in APPENDIX P) of the Newtonian Gravity Concept Inventory (NGCI), the first instrument to reliably measure Astro 101 student understanding of Newtonian gravity. The Construction phase involved writing clear items that minimize jargon and reading load. As these items were piloted, more information was gained about student understanding of gravity, which helped to refine the conceptual domain (Table 3). Performing a quantitative analysis using Classical Test Theory helped to show how well the construction of questions succeeded in producing items of appropriate difficulty and good discriminatory power. With the additional information from student interviews (audio recordings in APPENDIX C, and notes in APPENDIX N, APPENDIX R and APPENDIX S) and student response patterns (APPENDIX J, APPENDIX M, and APPENDIX Q), poorly functioning items were removed or revised in an iterative fashion. This process was discussed in the context of each of the four conceptual domains to show how the scope of these domains evolved and to show that the intended focus of the NGCI manifests in specific items. Finally, the Cronbach's alpha statistic of CTT showed the NGCI is a reliable instrument, and an argument-based approach was used to show that the assumptions one makes about interpreting NGCI scores as a measure of student understanding of Newtonian gravity are valid.

While the NGCI was developed and validated for the Astro 101 population, the topic of Newtonian gravity is also taught in introductory algebra-based college physics (Phys 101), courses, but in the context of very different curriculum and with a very different type of student body. Chapter 5 describes additional NGCI data from Phys 101 classes, calculates additional descriptive and CTT statistics, and presents additional arguments grounded in student interviews to show that the NGCI is indeed a reliable and valid instrument for also probing physic students' understanding of Newtonian gravity. Furthermore, this dual-purpose, extended applicability of the NGCI allows for comparisons between Astro 101 and Phys 101 students. Namely, it is shown that physics students have an overall stronger understanding of Newtonian gravity. For example, physics students had higher NGCI scores than astronomy students both pre- and post-instruction, physics CTT difficulty values were consistently lower than astronomy CTT values, and interviews with physics students suggest they have a greater overall understanding of Newtonian gravity. By comparing physics and astronomy CTT statistics and carefully analyzing the functioning of individual items, one finds that both astronomy and physics students hold several of the same conceptual and reasoning difficulties; however, there are subtle differences in how physics students approach each of the four NGCI conceptual domains. Physics students show a differentially higher comfort level with ground-based scenarios, where the gravitational force is "perpendicular to the surface." And, physics students have a slightly higher propensity to confuse rotational forces with gravitational force. However, they show greater scientific thinking, as evidenced by their greater use of proportional reasoning and understanding of the universality of gravity within certain thresholds.

To *calibrate* the NGCI and quantitatively *measure* students' learning of Newtonian gravity, Chapter 6 applies a Two-Parameter Logistic Item Response Theory (2PL IRT) Model to the dichotomously scored composite NGCI data consisting of the four-group (pre-instruction and post-instruction physics and astronomy) sample In order to perform the IRT analysis, the two assumptions of unidimensionality and local independence had to be met. A correlation and eigenvalue analysis of student IRT abilities and their responses to items showed that that the NGCI is sufficiently unidimensional, but Yen's Q3 statistic indicated that four items (Items 7, 10, 14, and 17) needed to be removed for the remaining 22 items to meet the local independence requirement. Fit statistics and item characteristic curves show that the resulting 2PL IRT model fits the data, meaning that parameters are *invariant*. Item parameters can be taken as inherent properties of the items (i.e. independent from the sample population, thereby calibrating the NGCI for all Astro 101 and Phys 101 classes), and student abilities estimates can be taken as inherent properties of the students (i.e. separate from the NGCI, thereby allowing for a quantitative analysis of student learning of Newtonian gravity).

The difference between an individual student's pre- and post-instruction IRT ability represents her gain, or learning, of Newtonian gravity. The different impact of physics compared to astronomy curriculum (i.e. topical context) on students' IRT gains was explored with a subset of 490 Astro 101 and 653 Phys 101 matched pre- and post-instruction student data via multiple linear regression models. With the dependent variable of IRT gain, the main effect variable of Curriculum is explored by controlling for seven other predictor variables, including four student characteristics variables (Pre-Instruction Ability, Major, Gender, and Previous Courses taken) and three classroom environment variables (professors number of Years Teaching, Institution Type, and level of Interactivity in the classroom). Results showed that the most significant factors in determining a student's learning of Newtonian gravity are Pre-Instruction Ability (predicting that students of higher ability will have lower gains, indicating a law of diminishing returns) and Interactivity Score (predicting that more interactive engagement classroom learning opportunities can lead to greater gains, indicating that instructors can enhance the impact of their class by providing more non-traditional learning opportunities). Major and Gender were also significant. However, the full regression model showed that the different effect of the discipline-specific topical focuses in physics versus astronomy courses was *not* a significant predictor of student learning of Newtonian gravity. While there are certainly other variables that could influence students' learning of Newtonian gravity that, when controlled, might allow some significant impact of Curriculum to emerge, the results of the regression models implemented in this project indicate that the topical context is not nearly as important as *how* an instructor engages students through interactivity in the classroom environment.

Overall, this research provides an important, robustly calibrated, reliable and valid tool, the NGCI, for measuring the effects of instruction on student learning of Newtonian gravity and for conducting important research on student difficulties when learning about Newtonian gravity. This tool exceeds the applicability of typical Concept Inventories in that its conceptual focus lends itself to measuring student learning in two very different contexts, i.e. Astro 101 and the Phys 101 courses. The NGCI is ready for widespread use and poised to lead to great strides in designing curricula and practices that maximize conceptual learning of Newtonian gravity across the United States.

Future Work

Perhaps one of the most important outcomes of this work is that it motivates a discussion of how Newtonian gravity should be taught in Astro 101 and Phys 101 so as to maximize learning. First, to assess learning on a local scale, individual instructors can and should use the NGCI for assessment in their classrooms. Currently, instructors can get a “first-order” estimate of their students’ performance by calculating their students’ Hake (1998) gains and the Classical Test Theory statistics for their sample. However, the NGCI and other Concept Inventories will have the most impact if instructors can compare their class data to an IRT calibrated database. The person-independent Item Response Theory item parameters put forth in this dissertation should be used to estimate students’ true abilities along the same scale, i.e. along the linear Newtonian gravity construct. While it would be difficult for an instructor to implement an IRT analysis alone, an online, calibrated database that could implement the analysis “behind the scenes” could streamline and standardize the assessment process, providing rapid and robust information about student learning within the broader context.

On a wider scale, 2PL IRT calibrated abilities for students from a much larger data set from Astro 101 and Phys 101 classes at many more institutions could help to establish national norms on student understanding and learning of Newtonian gravity. Researchers can compare pre- and post-instruction ability values for different instructional interventions to find which methods of instruction lead to the greatest Newtonian gravity IRT abilities and pre- to post-instruction gains. Kavanagh and Sneider (2007a and 2007b) suggest that instructional interventions at the K-12 level will have the most impact if they provide the same cognitive dissonance that scholars grappled with historically. However, the evidence from research on student learning at the college level (Hake 1998, Prather et al 2009, Rudolph et al 2010) and from the regression analysis of Chapter 6 suggests that for the Astro 101 and Phys 101 populations, interactivity in the classroom is the primary classroom environment variable in leading to conceptual change. What kinds of interactivity methods work the best? What is the differential effect of laboratories, tutorials, Think-Pair-Share questions, and Ranking Tasks (Hudgins et al 2006) on student learning of Newtonian gravity? The *Lecture-Tutorials for Introductory Astronomy* (Prather et al 2012) include a tutorial on Newtonian gravity. How could it be improved to address the student reasoning difficulties uncovered and discussed in this dissertation? Addressing these questions will give Astro 101 and Phys 101 instructors the capacity to more effectively reach a broad range of students.

The avenues of future research discussed above can be accomplished by coding student responses dichotomously and interpreting student scores and abilities. However, because the NGCI was created with robust information about student reasoning difficulties, distractor choices offer rich information about specific student misconceptions that can be linked to specific mental models, i.e. those discussed in Chapter 3. While Chapters 4 and 5 explored student response patterns to get an idea of the prevalence of students choosing some distractors (i.e. using some mental models over others), a more systematic, quantitative approach could be undertaken with Model Analysis (Bao 1999, Bao & Redish 2006, Ding & Beichner 2009). Model Analysis leverages the mathematical framework of quantum mechanics to account for the probabilistic nature of the interaction of a student's mental landscape with survey questions. A "model state" may represent a correct scientific model, a documented alternative mental model or misconception, or some other naïve idea, and the set of these models forms an orthogonal basis. A linear vector can then be defined for each student, with a coefficient for each basis state that represents the probability of using different models in answering a set of questions that probe the same idea but have different surface features. Using this student state vector to create a "density matrix" can quantitatively show how consistently a student uses each mental model. Analyzing the Astro 101 and Phys 101 student response data with Model Analysis could more definitively characterize the observed differences between physics and astronomy students' approaches to questions within each of the four NGCI conceptual domains (Table 3).

One final possible avenue of research on student understanding of gravity would be to broaden the conceptual domain to the Einsteinian perspective. While adopting a strictly Newtonian perspective was necessary for achieving a clear and manageable focus for this research project, including both perspectives would be more holistic. After all, novice students often do not maintain or even understand a clear distinction between Newtonian and Einsteinian gravity (the word "Newtonian" was included in the title of the NGCI for the sake of the experts taking the survey). Before ever taking an astronomy or physics course, students learn about gravity from daily experience and from popular culture, which includes some mixture of both theories. Furthermore, Einstein's theory accounts for the shortcomings of Newton's theory, grounding such an extension in the historical development of a unified theory of gravity.

REFERENCES CITED

Agan, L. and Sneider, C. 2004, "Learning about the Earth's Shape and Gravity: A Guide for teachers and curriculum developer," *Astronomy Education Review*, 2, 90.

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

Andrich, D. 2004, "Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms?" *Medical Care*, 42(1), 17.

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

Aubrecht, G. J, and Aubrecht, J. D. 1983, "Constructing Objective Tests," *American Journal of Physics*, 51, 613.

Bailey, J. 2006, "Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties about the Properties and Formation of Stars," Doctoral Thesis in Teaching and Education, University of Arizona.

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

Bailey, J. M., Prather, E. E., Johnson, B., and Slater, T. F. 2009, "College Students' Preinstructional Ideas About Stars and Star Formation," *Astronomy Education Review*, 8, 010110-1.

Bailey, J. M., Johnson, B., Prather, E. E., and Slater, T. F. 2011, "Development and Validation of the Star Properties Concept Inventory," *International Journal of Science Education*, 1.

Bao, L. 1999, "Dynamics of Student Modeling: A Theory, Algorithms, and Application to Quantum Mechanics," Doctoral Thesis in Physics, University of Maryland.

Bao, L. and Redish, E. F. 2006, "Model analysis: Representing and assessing the dynamics of student learning," *Physical Review Special Topics – Physics Education Research*, 2, 1.

Baker, F. B. and Kim, S. 2004, Item Response Theory: Parameter Estimation Techniques, 2nd ed., New York: Dekker.

Bar, V., Zinn, B., Goldmuntz, R., and Sneider, C. 1994, "Children's Concepts about Weight and Free Fall," *Science Education*, 78(2), 149.

Bar, V., Zinn, B., and Rubin, E. 1997, "Children's Ideas About Action at a Distance," *International Journal of Science Education*, 19(10), 1137.

Bar, V., and Zinn, B. 1998, "Similar Frameworks of Action-at-a-Distance: Early Scientists' and Pupils' Ideas," *Science Education*, 7(5), 471.

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.

Bliss, J., Ogborn, J., and Whitelock, D. 1989, "Secondary School Pupils' Commonsense Theories of Motion," *International Journal of Science Education*, 11(3), 261.

Bock, R. D. and Lieberman, M. 1970, "Fitting a response model for *n* dichotomously scored items," *Psychometrika*, 35, 179-197.

Bock, R. D. and Aitkin, M. 1981, "Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm," *Psychometrika, 46, 443-459.*

Bodner, G. M. 1986, "Constructivism: A Theory of Knowledge," *Journal of Chemical Education*, 63(10), 873.

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: Methodology for Determining Cognitive and Communicative Processes in Survey Research, San Francisco: Jossey-Bass.

Bond, T. G., and Fox, C. M. 2001, Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Mahwah, NJ: Lawrence Erlbaum Associates.

Brace, N., Kemp, R., and Snelgar, R. 2013, SPSS for Psychologists, 5th ed., New York: Macmillan Publishers Ltd.

Bransford, J. D., Brown, A. L., and Cocking, R. R. (Eds.) 1999, How people learn: Brain, Mind, Experience, and School. Washington, DC: National Academy of Sciences.

Caramazza, A., McCloskey, M., and Green, B. 1983, "Naïve Beliefs in 'Sophisticated' Subject Misconceptions about Trajectories and Objects," *Cognition*, 9(2), 117.

Champagne, A., Klopfer, L., and Anderson, J. 1980, "Factors Influencing the Learning of

Classical Mechanics," *American Journal of Physics*, 48(12), 1074.

Chandler, D. 1991, "Weightlessness and Microgravity," *Physics Teacher*, 29, 312.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. 1981, "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science*, 5(2), 121.

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.

Clement, J. 1982, "Students' Preconceptions in Introductory Mechanics," *American Journal of Physics*, 50(1), 66.

Cohen, I. B. and Smith, G. E. (Editors) 2002, The Cambridge Companion to Newton, Cambridge, United Kingdom: Cambridge University Press.

Cohen, J. 1960, "A Coefficient of Agreement for Nominal Scales," *Education and Psychological Measurement*, 20(1), 37.

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.

Cronbach, L. J. 1951, "Coefficient alpha and the internal structure of tests," *Psychometrika*, 16, 297.

de Ayala, R. J. 2009, The Theory and Practice of Item Response Theory, New York: The Guilford Press.

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

diSessa, A. 1993, "Toward an Epistemology of Physics," *Cognition and Instruction*, 10, 105.

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

Driver, R., Asoko, H., Leach, J., Scott, P., and Mortimer, E. 1994, "Constructing Scientific Knowledge in the Classroom," *Educational Researcher*, 23, 5.

Fan, X. 1998, "Item Response Theory and Classical Test Theory: an Empirical Comparison of their Item/Person Statistics," *Educational and Psychological Measurement*, 58(3), 357.

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

Fetter, A. L. and Walecka, J. D. 2003. Theoretical Mechanics of Particles and Continua, Mineola: Dover Publications, Inc.

Field, A. 2005. Discovering Statistics Using SPSS. 2nd ed. London: Sage.

Fosnot, C. T. (Ed.) 1996. Constructivism: Theory, Perspectives, and Practice, New York: Teachers College Press.

Franknoi, A. 2002, "Enrollments in Astronomy 101 Courses," *Astronomy Education Review*, 1(1), 121.

Galili, I. 1995, "Interpretation of Students' Understanding of the Concept of Weightlessness," *Research in Science Education*, 25(1), 51.

Galili, I. 2001, "Weight Versus Gravitational Force: Historical and Education Perspectives," *International Journal of Science Education*, 23, 1073.

Galili, I. and Lehavi, Y. 2003, "The Importance of Weightlessness and Tides in Teaching Gravitation," *American Journal of Physics*, 71(11), 1127.

Graham, T. and Berry, J. 1993, "Students' intuitive understanding of gravity," *International Journal of Mathematical Education in Science and Technology*, 243, 473.

Griffiths, D. J.1999. Introduction to Electrodynamics, 3rd Ed. Upper Saddle River: Prentice Hall.

Gunstone, R. F. and White, R. T. 1980, "A matter of gravity," *Research in Science Education*, 10, 35.

Gunstone, R. F. and White, R. T. 1981, "Understanding of gravity," *Science Education*, 65, 291.

Gunstone, R. F. and Watts, M. 1985, "Force and Motion," in *Children's Ideas in Science*, Driver, R., Guesne, E., and Tiberghien, A. (Editors), Milton Keynes, UK: Open University Press, 85.

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.

Hake, R. R. 1998, "Interactive-Engagement Versus Traditional Methods: A Six-thousand-student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, 66(1), 64.

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.

Halloun, I. and Hestenes, D. 1985a, "The Initial Knowledge State of College Physics Students," *American Journal of Physics*, 53(11), 1043.

Halloun, I. and Hestenes, D. 1985b, "Common Sense Concepts about Motion," *American Journal of Physics*, 53(11), 1056.

Hambleton, R. K. and Jones, R. J. 1993, "Comparison of Classical Test Theory and Item Response Theory and Their Application to Test Development," *Educational Measurement: Issues and Practice*, 12, 38.

Hambleton, R. K. and Swaminathan, H. 1985. Item Response Theory: Principles and Applications, Kluwer-Nijhoff Publishing, Boston/Dordrecht/Lancaster.

Harris, D. 1989, "Comparison of 1-, 2-, and 3-Parameter IRT Models," *Educational Measurement: Issues and Practice*, 8, 35.

Harwell, M. R. and Baker, F. B 1991, "The Use of Prior Distributions in Marginalized Bayesian Item Parameter Estimation: A Didactic," *Applied Psychological Measurement*, 15, 375.

Hestenes, D., Wells, M., and Swackhamer, G. 1992, "Force Concept Inventory," *The Physics Teacher*, 30, 141.

Hobson, A. 2008, "The Surprising Effectiveness of College Scientific Literacy Courses," *The Physics Teacher*, 46, 404.

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.

Jones, B. L., Lynch, P. P., and Reesink, C. 2007, "Children's Conceptions of the Earth, Sun and Moon," *International Journal of Science Education*, 9(1), 43.

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

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

Kavanagh, C. and Sneider, C. 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 thesis, University of Arizona.

Landis, J. R. and Koch, G. G. 1977, "The measurements of observer agreement for categorical data," *Biometric*, 33, 159.

Lee, Y., Palazzo, D. J., Warnakulasooriya, R., and Pritchard, D. E. 2008, "Measuring Student Learning with Item Response Theory," *Physics Review Special Topics: Physics Education Research*, 4, 10102.

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

Lord, F. M. and Novick, M. R. 1968, Statistical Theories of Mental Test Scores, Reading, MA: Addison-Wesley.

Mali, G. and Howe, A. 1979, "Development of Earth and Gravity Concepts among Nepali Children," *Science Education*, 63(5), 685.

Maloney, D., O'Kuma, T., Hieggelke, C., and Van Heuvelen, A. 2001, "Surveying Students' Conceptual Knowledge of Electricity and Magnetism," *American Journal of Physics*, 69, S12.

McCloskey, M. 1983, "Naïve Theories of Motion," in Mental Models, Gentner, D. and Stevens, A. (Editors), Hillsdale, NJ: Erlbaum, 299.

McCloskey, M., Washburn, A., and Felch, L. 1983, "Intuitive Physics: The Straight-Down Belief and Its Origin," *Journal of Experimental Psychology: Learning, Memory & Cognition*, 9(4), 636.

McDermott, L. C., 1991, "Millikan Lecture 1990: What We Teach and What is Learned – Closing the Gap," *American Journal of Physics*, 59, 301.

McDermott, L. C. and the Physics Education Group at the University of Washington 1996, Physics by Inquiry, New York: John Wiley & Sons.

McDermott, L. C., Shaffer, P. S., and the Physics Education Group at the University of Washington 2002, Tutorials in Introductory Physics, Pearson Publishing, Inc.

Meltzer, D. E. and Manivannan, K. 1996, "Promoting Interactivity in Physics Lecture Classes," *The Physics Teacher*, 34(2), 72.

Minstrell, J. A. 1989, "Teaching Science for Understanding," in Toward the Thinking Curriculum: Current and Cognitive Research, Resnick, L. B. and Klopfer, L. E. (Editors), Alexandria, VA: Association for Supervision and Curriculum Development, 129.

Mislevy, R. J. 2006, "Cognitive Psychology and Educational Assessment," in Educational Measurement, Brennan, R. L. (Ed.) Westport, CT: Praeger Publishers.

Newton, I. 1967, translation by Cohen, I. B., Whitman, A. with assistance by Budenz, J., The Principia: Mathematical Principles of Natural Philosophy, Berkeley, CA: University of California Press.

NGSS Lead States 2013, Next Generation Science Standards, Washington, DC: The National Academies Press.

Noce, G., Torosantucci, G., and Vincentini, M. 1988, "The Floating of Objects on the Moon: Prediction from a Theory or Experimental Facts?" *International Journal of Science Education*, 10(1), 61.

Nussbaum, J. and Novak, J. 1976, "An assessment of children's concepts of earth utilizing structured interviews," *Science Education*, 60, 535.

Nussbaum, J. 1979, "Children's Conceptions of the Earth as a Cosmic Body: A Cross Age Study," *Science Education*, 63(1) 83.

O'Brien, R. M. 2007, "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Quality and Quantity*, 41, 5, 673.

Ogborn, J. 1985, "Understanding Students' Understanding: An Example from Dynamics," *European Journal of Science Education*, 7(2), 141.

Osborn, R. 1983, “Building on Children’s Intuitive Ideas,” in R. Osborne & P. Freyberg (Editors), Learning in Science: the Implications of Children’s Science, Portsmouth, NH: Heinemann, 41.

Palmer, D. 2001, “Students’ alternative conceptions and scientifically acceptable conceptions about gravity,” *International Journal of Science Education*, 23, 691.

Partridge, P. and Greenstein, G. 2003, “Goals for Astro 101: Report on Workshop for Department Leaders,” *Astronomy Education Review*, 2, 46. Paper presented at the 61st annual meeting of the national association for research in science teaching,

Piburn, M. 1988, “Misconceptions about Gravity Held by College Students,” conference proceedings, NARST Annual Meeting, Lake of the Ozarks, Missouri. April 10-13. (ERIC Document Reproduction Service No. 292 616).

Planinic, M., Ivanjek, L., and Susac, A. 2010, “Rasch Model Based Analysis of the Force Concept Inventory,” *Physics Review Special Topics: Physics Education Research*, 6, 010103.

Piaget, J. 1952, The Origins of Intelligence in Children, translated by M. Cook, New York: International Universities Press.

Prather, E. E., Rudolph, A. L., Brissenden G. and Schlingman, W. M. 2009, “A National Study Assessing the Teaching and Learning of Introductory Astronomy. Part I. The Effect of Interactive Instruction,” *American Journal of Physics*, 77, 320.

Prather, E. E., Rudolph, A. L., and Brissenden, G. 2009, “Teaching and learning astronomy in the 21st century,” *Physics Today*, 62, 41.

Prather, E. E., Slater, T. F., Adams, J. P., and Brissenden, G. 2012, Lecture-Tutorials for Introductory Astronomy (3rd ed.). San Francisco, CA: Pearson Addison-Wesley.

Redish, E. F. 1994, “The Implications of Cognitive Studies for Teaching Physics,” *American Journal of Physics*, 62, 796.

Redish, E. F. 2003, Teaching Physics with the Physics Suite, New York: J. Wiley.

Reynoso, E., Fierro, E., and Torres, G. 1993, “The Alternative Frameworks Presented by Mexican Students and Teachers Concerning the Free Fall of Bodies,” *International Journal of Science Education*, 15(2), 127.

Rodriguez, I., Brewe, E., Sawtelle, V., and Kramer, L. H. 2012, “Impact of Equity Models and Statistical Measures on Interpretations of Educational Reform,” *Physical Review Special Topics – Physics Education Research*, 8, 020103.

Rudolph, A. L., Prather, E. E., Brissenden, G., Consiglio, D., and Gonzaga, V. 2010, "A National Study Assessing the Teaching and Learning of Introductory Astronomy Part II: The Connection between Student Demographics and Learning," *Astronomy Education Review*, 9, 010107-1.

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.

Rupp, A. A. and Zumbo, B. D. 2006, "Understanding Parameter Invariance in Unidimensional IRT Models," *Education and Psychological Measurement*, 66, 63.

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.

Sequeira, M., and Leite, L. 1991, "Alternative Conceptions and History of Science in Physics Teacher Education," *Science Education*, 75(1), 45.

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.

Slater, T. F., Adams, J. P., Brissenden, G., and Duncan, D. 2001, "What Topics Are Taught in Introductory Astronomy Courses?" *The Physics Teacher*, 39, 52.

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.

Sneider, C. and Pulos, S. 1983, "Children's Cosmographies: Understanding the Earth Shape and Gravity," *Science Education*, 67(2), 205.

Taylor, J. R. 1982, <u>An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements</u>. Mill Valley, CA: University Science Books.

Treagust, D., and Smith, C. 1989, "Secondary Students' Understanding of Gravity and the Motion of Planets," *School Science and Mathematics*, 89(5), 380.

Urdan, T. C. 2010, <u>Statistics in Plain English</u>, 3rd ed. New York: Taylor & Francis Group.

Viennot, L. 1979, "Spontaneous Reasoning in Elementary Dynamics," *European Journal of Science Education*, 1(2), 205.

Viennot, L. 1985, "Analysing students' reasoning in science: A pragmatic view of theoretical problem," *European Journal of Science Education*, 7(2), 151.

von Glaserfeld, E. 1981 in "New Directions in Piagetian Theory and Practice," Sigel, I. E., Bordzinsky, D. M., Golinkoff, R. M. (Editors), Hillside, NJ: Erlbaum.

Vosniadou, S. and Brewer, W. F. 1987, "Theories of Knowledge Restructuring in Development," *Review of Educational Research*, 57, 51.

Vosniadou, S. 1994, "Capturing and Modeling the Process of Conceptual Change," *Learning and Instruction*, 4, 45.

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

Wallace, C. 2011, "An Investigation into Introductory Astronomy Students' Difficulties with Cosmology, and the Development, Validation, and Efficacy of a New Suite of Cosmology Lecture-Tutorials," Doctoral Thesis in Astrophysical and Planetary Sciences, University of Colorado.

Wang, J. and Bao, L. 2010, "Analyzing Force Concept Inventory with Item Response Theory," *American Journal of Physics*, 78(10), 1064.

Watts, D.M. 1982, "Gravity – Don't Take it for Granted!" *Physics Education*, 17(4), 116.

Whitelock, D. 1991, "Investigating a Model of Commonsense Thinking about Causes of Motion with 7-16 Year Old Pupils," *International Journal of Science Education*, 13(3), 321.

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.

Wu, M. and Adams, R. J. (unpublished), "Properties of Rasch Residual Fit Statistics," *Journal of Applied Measurement*.

Yen, W. M. 1984, "Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model," *Applied Psychological Measurement*, 8, 125.

Yen, W. M. and Fitzpatrick, A. R. 2006, "Item Response Theory," in *Educational Measurement*, 4th Ed., Brennan, R. (Editor), Westport, CT: American Council on Education/Praeger, 111-153.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. 1996, BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items, Chicago, IL: Scientific Software.

APPENDICES

APPENDIX A

EXPERT GRAVITY INTERVIEW TEMPLATE

Expert Gravity Interview

Name:
Status:
Date/Time:

1. What is gravity?

2. What does the general public think gravity is?

3. What *should* a college student understand about gravity?

    a. What *do* they understand?

4. How important is it that people understand gravity?

5. Does learning about gravity offer a good opportunity to think scientifically?

APPENDIX B

GENERAL INTERVIEW QUESTIONNAIRE ABOUT GRAVITY

Student Gravity Interview

Student ID:
Major:
Year:
Date:

Note: The following questions apply to the Newtonian picture of gravity.

1. What is gravity?

2. What causes gravity?

3. How could you change the strength of gravity that an object experiences?

4. Can you think of a place where there is no gravity?

5. Does gravity have a direction?

6. How do the tides work?

7. Why are planets and stars round, but other objects like asteroids are not?

8. You have probably heard the story of an apple falling on Newton's head, and then he suddenly realizes that the thing that makes the apple fall to the Earth is the same thing that makes the Moon orbit the Earth and the Earth orbit the Sun. How can this be?

9. Can you write down an equation that describes gravity?

10. Why is gravity important?

APPENDIX C

AUDIO-RECORDED INTERVIEWS

Audio recordings are organized according to the following table. To request CD copies contact your local public or university library to place and interlibrary loan request to Montana State University. Questions call 406-994-3161.

Semester	Topic	Interviewees	Number of Interviews Conducted
Spring 2010	Expert Gravity Interview Template (APPENDIX A)	Physics, astronomy, and geoscience experts	6
Spring 2010	General Gravity Questionnaire (APPENDIX B)	Astro 101 students	18
Fall 2010	Open Response Questionnaires (APPENDIX E, APPENDIX F, and APPENDIX G)	Astro 101 students	15
Fall 2011	Gravity Concept Inventory Version 2 (APPENDIX L)	Astro 101 students	17
Spring 2012	Newtonian Gravity Concept Inventory Version 3 (APPENDIX P)	Astro 101 students	7
Fall 2012	Newtonian Gravity Concept Inventory Version 3 (APPENDIX P)	Phys 101 students	11

APPENDIX D

STUDENT INTERVIEW NOTES FOR GENERAL GRAVITY QUESTIONNAIRE

Student Information	Interview Notes
S10-01 Freshman Environmental Science	- Gravity keeps people down. It's caused by the core of Earth, or any large body. Every particle has gravity. Could change strength by changing the electric and magnetic forces around the object, although not sure. Maybe proximity of other forces and alignment of Moon and Sun. Points to the largest, density object within proximity. - Stars/planets are round because they rotate faster, cool slower, and are less dense. More material can migrate, easier to shape. - Equation is a balance between pull and density and centrifugal force. - Air pressure depends on gravity. Otherwise particles wouldn't bind together.
S10-02 Freshman Civil Engineering	- Gravity is the pull of something of an object of equal or lesser size. Keeps stuff on the ground. Caused by the amount of mass. Could change with more volume or pressure. In a vacuum there is more gravity. - There is some gravity in space. Less on the Moon. Space station has artificial gravity. - Planets/stars are round because of gravitational pull, asteroids have broken off. - Moon doesn't fall because it is too big. It has polarity. Gravity pushes in some places and pulls in others, like a magnet. - Equation would be a downward force plus the size of the object. - Gravity keeps everything from flying away.
S10-03 Junior Film	- Gravity is a force that makes us rotate around the Sun, keeps everything grounded. Not sure what causes it, maybe large mass, so could change strength by changing mass. Direction on Earth is toward core, and between Earth and Sun it's toward the Sun. - No gravity in absolute space, outside solar systems. There is less gravity on the Moon, so a pen would float. Astronauts could stand because they had larger mass. - Asteroids have broken off from something else. Stars/planets were born spherical. - Moon doesn't fall because the gravity between the Moon and Earth isn't enough. They aren't close enough. Apple falls because it's in the atmosphere. - G, F, mass of the object orbiting. - Gravity has to do with how far we are from the Sun, the orbit

		is just where our planet fell into.
S10-04 Sophomore Geology	- - - - -	Gravity is a force pulling two masses toward each other. Caused by atoms and electron attraction. Could change strength by putting it next to a larger mass, or accelerate it, but that may be an illusion. You would just feel heavier or lighter. Direction is toward center of some object that has mass. There is gravity in space. Moon doesn't fall because it is farther or has enough speed so that keeps it moving. Equation involves mass of each object. Gravity holds the universe together.
S10-05 Freshman Secondary Education	- - - - -	Gravity is an inherent property of matter, it keeps us from flying off Earth. Don't know what causes it, but everything has it, even us. Could change strength by changing matter (density, size, mass, material), put in a vacuum, push under water, basically anywhere there are other forces. Direction is toward center of an object. Applies to all shapes. Tides are because the Moon's interaction is an ellipse. Planets/stars are spherical and asteroids are not because they are not formed the same way. Other forces prevent Moon from colliding, probably magnetism. Apples don't have magnetic fields. Like when stars dies some forces pull in and others push out. Equation is mass divided by something, maybe magnetic term.
S10-06 Freshman Civil Engineering	- - - - - -	Gravity is the force that pulls down toward the center of Earth. Not sure if it's created by magnetic field in Earth's core or mass. Can't change the strength, but you can have force to counteract it. Less if you go to space or dive down to feel weightless. Direction points down, toward Sun and Earth. Weaker at edge of atmosphere. No gravity in space as far from stars as possible. Stars/planets are active/alive. Asteroids are dead planets. Moon doesn't fall because it is fast enough. a = 0.8 m/s^2. Air resistance also matters somehow. Gravity prevents floating.
S10-07 Freshman Business Marketing	- -	Gravity is a force that keeps things grounded, it's always consistent, caused by forces in the atmosphere, could be changed by adding weight or strengthening magnetic fields. Always thought that there is no gravity in space, but was told that there is – doesn't understand. It keeps the planets from running around, keeps them in their orbit.

	- Gravity is down on Earth, but has no direction in space. - Moon doesn't fall down because it's not in our atmosphere. - F = mc, g = ma, but maybe that's just on Earth.
S10-08 Freshman Undeclared	- Gravity is mass being attracted to each other, like magnets pulling. Toward center/middle; it all wants to go to the same place. No idea what causes gravity, haven't thought about it before. - Gravity is everywhere, maybe it would be zero if you're outside the boundary (gradual). The Moon doesn't fall because Earth's gravity isn't strong enough. The Moon is too far away, but it's outside the boundary, it's at the boundary. - Sun's gravity is holding planets. - mc^2, mass, force = gravity - Gravity holds our atmosphere in, makes formation of anything in the universe possible.
S10-09 Junior Applied Math	- Force of gravity is 9.8 m/s^2 on Earth's surface, different elsewhere. - Don't know what causes gravity; maybe orbit of Earth around Sun. Could change strength by moving up in elevation or maybe in water. - No gravity when weightless (space, references Vomit Comet) - Moon doesn't fall to Earth like an apple does because it is farther away. - F = mg
S10-10 Freshman Liberal Studies – Environmental focus	- Gravity is a force exerted by anything with mass. Caused by mass. Could change strength by changing distance from other objects, speed they're passing by, mountains change a little. - Gravity does not have a direction but has vectors. - Gravity is close to zero in space. - The Moon doesn't fall because it has speed to counteract gravity. - Equation is some vector relation between mass and velocity. - Pen on Moon would fall toward center of mass, but at a slower rate.
S10-11 Senior Political Science	- Gravity is one of the four forces; it is the attraction of masses to each other. Caused by curved spacetime. Can change strength by adding/subtracting mass, or changing distance. - Maybe there is somewhere in space that is void of mass, so void of gravity. - Not sure why there is a tidal bulge on the side of Earth opposite the Moon. Probably Earth's electromagnetic field, maybe the Moon's gravity isn't strong enough that far away.

	- Asteroids break off from other collections of mass. They don't go through a formation stage. The longer it forms, the more likely it is to be spherical. - Moon doesn't fall because it's far enough away; it doesn't cross the tipping point. - F = GMm/r^2
S10-12 Freshman Music Education	- Gravity pulls you to Earth because Earth is so massive. - May be no gravity outside of Earth, but space station orbits, so there must be some. - Stars/planets are round because so much compression. Asteroids are broken off from planets. - Moon doesn't fall because it's revolving, so it just stays in that path.
S10-13 Non-traditional student Civil Engineering	- Gravity is an attraction between two massive bodies; they are pulled toward each other. Not sure what causes gravity. - Direction of gravity is toward the center of mass (for example, of Earth and the Moon in both directions). - When objects reach a certain mass they become round. Asteroids' gravity isn't strong enough; must be Pluto-sized or so. - Moon doesn't fall down because it is constantly falling, like a canon, it has enough velocity. But, the Moon is moving away from Earth so that doesn't make sense… maybe it's just far enough away? - Equation involves mass, distance, velocity, maybe relating to escape velocity.
S10-14 Sophomore Music Education and Spanish	- Gravity is the force that holds us on Earth, keeps planets aligned. No idea what causes it. Can change by making object bigger or closer to source. Has no direction. - No gravity in space. Space station has artificial gravity to keep it orbiting. There are gravitational forces but you don't experience them. - Planets have their own gravity, but asteroids don't because they are made of something else. - Force isn't falling, it's pulling. It keeps the Moon where it wants it. Gravity pulls and pushes because Moon has its own gravity and an apple doesn't. - Equation includes force, acceleration, time (how long it takes to fall), mass. Air affects gravity too. - Gravity affects everything about the way we live.
S10-15 Freshman Architecture	- Gravity is a force of attraction between two bodies, relative to their mass. Don't know what causes it. Would be changed only by changing mass, nothing else. It points to the center of a body.

	-	No gravity in space, but astronauts do experience gravity because they are orbiting.
	-	Tidal bulge on far side is from a disruption/rebound.
	-	Maybe asteroids are made of less dense materials so their gravity is less.
	-	Moon doesn't fall because it is going fast enough, like water in a bucket.
	-	m_1m_2, d_1d_2, there is a square somewhere.
	-	Gravity is incredibly important. Otherwise there would be no formation of stars/planets/life.
S10-16 Senior Business Management/ Marketing	-	Gravity is a force that keeps us on Earth. It's between two objects and is attractive. Caused by having mass and size, so could change strength by changing mass and distance (closer = more gravity).
	-	No place exists where there is zero gravity.
	-	Direction is to the center of the more massive object.
	-	Tidal bulge opposite Moon may be from the Sun or other planets.
	-	Asteroids rotate so they don't have strong enough gravity. A pen wouldn't have its own gravity.
	-	The Moon is falling, but it's so far away that the distance change is small, but it will eventually crash into Earth.
	-	Equation involves mass, distance, time (rate at which it's falling), velocity.
	-	Gravity is important; it keeps us on Earth, allows life. People should understand the basics of gravity; it will be more important in the future.
S10-17 Freshman Psychology	-	Gravity is a force that keeps things together, like planets. It has to do with density of a planet, what we base weight off of. Mass and distance cause gravity, so changing them would cause a change in gravity.
	-	Gravity is everywhere and has no direction.
	-	Earth keeps the Moon in orbit, and it pushes.
	-	Moon doesn't fall because of our atmosphere, which is pushes it because of density and mass.
	-	F = GMm/r^2. Doesn't know what G means, maybe gravity. Sum the masses, r is the rotation rate, i.e. spinning faster means weaker gravity.
	-	A pen would float on the Moon.
S10-18 Junior Computer Science	-	Gravity is a force exerted by a massive body on other masses, depends on the distance. Mass, maybe gravitons cause gravity. Has direction relative to the object you are feeling. Change the strength by changing mass and distance. You are heavier on top of a mountain because there is more

	material between you and Earth. - Planets are round because they have more mass and the mass wants to go to the center. They are also more geologically active, which smooths the surface. - Moon doesn't fall because it formed with an escape velocity. It's still trying to fall. - F = GMm/d - Gravity is fundamental, universal. Important for interstellar travel and walking.

APPENDIX E

OPEN-RESPONSE FORM A WITH RESULTS

Questions about Gravity
Version A
Fall 2010

1. Why do astronauts appear to float in their spacecraft? (**κ**=0.84)

Response Categories	N	Percent of Total
a. There is no gravity in space.	a. 72	a. 50.3%
b. Gravity is much weaker in space.	b. 21	b. 14.6%
c. They are too far away from Earth or any massive body.	c. 15	c. 10.2%
d. They are in a constant state of free fall.	d. 14	d. 9.7%
e. The spacecraft's gravity isn't strong enough.	e. 3	e. 2.1%

2. If you traveled down a tunnel that reached half way to the center of the Earth, would the strength of gravity you experience from the Earth change? If so, how? Explain your answer. (**κ**=0.88)

Response Categories	N	Percent of Total
a. No change, Earth's gravity is constant and uniform.	a. 33	a. 23.1%
b. Increase, you are closer to the center	b. 19	b. 13.3%
c. Increase, there is more pressure.	c. 9	c. 6.3%
d. Increase, closest to the densest area.	d. 7	d. 4.9%
e. Decrease, more mass above you and less beneath you.	e. 5	e. 3.5%
f. Increase, closest to the source of gravity.	f. 5	f. 3.5%
g. References interaction with magnetic field	g. 4	g. 2.8%
h. Discusses influence from other celestial bodies	h. 4	h. 2.8%
i. No change, Earth's mass is the same.	i. 4	i. 2.8%
j. Increase, smaller radius from center.	j. 3	j. 2.1%

3. Is there a relationship between a planet's gravity and its atmosphere? If so, describe this relationship. (**κ**=0.85)

Response Categories	N	Percent of Total
a. No, not that they know of.	a. 29	a. 20.3%
b. Yes, gravity holds the atmosphere to the planet.	b. 23	b. 16.1%
c. Yes, atmosphere affects the strength of gravity.	c. 17	c. 11.9%
d. Yes, gravity determines the thickness and composition of the atmosphere.	d. 10	d. 7.0%
e. No, gravity depends on mass and distance and can exist without atmosphere.	e. 10	e. 7.0%

f. Yes, atmosphere represents the boundary of gravity.	f. 8	f. 5.6%
g. Yes, gravity causes the atmosphere to exist.	g. 6	g. 4.2%
h. Yes, the atmosphere exerts pressure.	h. 6	h. 4.2%
i. Yes, but not sure what relationship.	i. 5	i. 3.5%

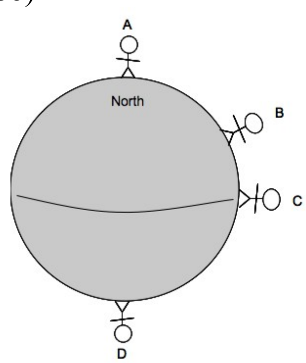
4. How heavy or light does something have to be to create its own gravitational field? (**κ**=0.93)

Response Categories	N	Percent of Total
a. Every object has some gravitational field.	a. 61	a. 42.7%
b. Very, very heavy.	b. 27	b. 18.9%
c. Planet/moon sized.	c. 10	c. 7.0%
d. Has gravity only if it has an effect on other objects.	d. 8	d. 5.6%
e. Not everything, at least needs to be tangible.	e. 5	e. 3.5%

5. During a Lunar Eclipse, the Earth is directly between the Sun and the Moon, causing it to cast a shadow on the Moon. How does the gravitational attraction between the Earth and the Moon change during this event (if at all)?

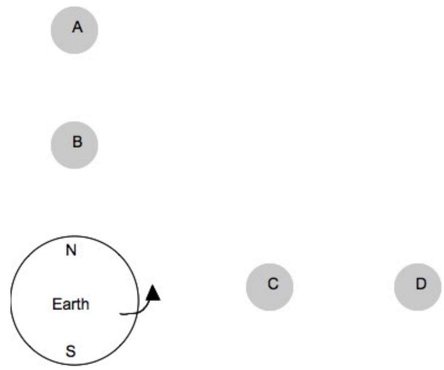
Student responses to this question did not provide any new, previously undocumented information.

6. Which person standing on the surface of the Earth experiences a stronger force of gravity? Why? (**κ**=0.86)



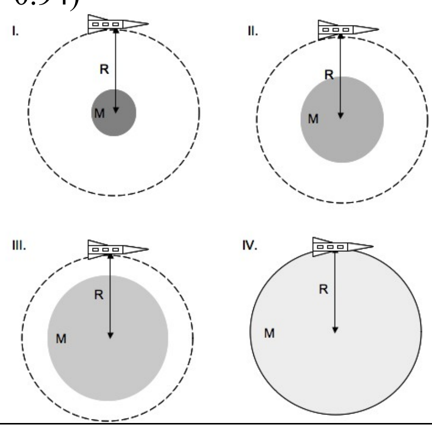
Response Categories	N	Percent of Total
a. Gravity is the same everywhere on the planet.	a. 73	a. 51.0%
b. The equator is special.	b. 13	b. 9.1%
c. Thinks that South is "down".	c. 8	c. 5.6%
d. Considers rotation.	d. 6	d. 4.2%
e. Accounted for the oblateness of Earth.	e. 4	e. 2.8%
f. Considers magnetism.	f. 4	f. 2.8%
g. Considers position of Sun/Moon.	g. 3	g. 2.1%

7. Describe how the strength of gravity <u>on Earth</u> compares for each possible position of the Moon.



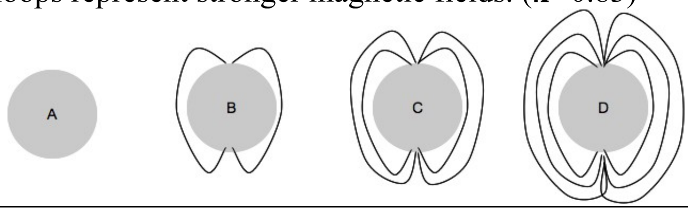
Student responses to this question did not provide any new, previously undocumented information.

8. Each of the following planets has the same mass M, but each is made of a different material. The darker the planet, the denser the material. Describe the strength of gravity that a rocket at distance R from the center of each planet experiences. Why? (**κ**=0.94)



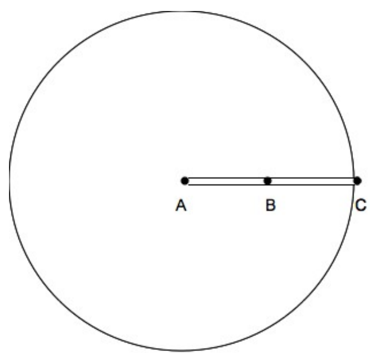
Response Categories	N	Percent of Total
a. Same, equal mass and distance from center.	a. 27	a. 18.9%
b. Denser material = stronger gravity.	b. 18	b. 12.6%
c. Must be "just right" – planet not too dense and rocket not too far (chose B or C).	c. 17	c. 11.9%
d. A is the strongest and D is weakest, no explanation.	d. 15	d. 10.5%
e. Considers distance from the surface of the planet.	e. 11	e. 7.7%
f. A is the weakest and D is strongest, no explanation.	f. 8	f. 5.6%
g. All same, higher density is compensated by smaller distance to surface.	g. 6	g. 4.2%
h. The size of the planet determines gravity.	h. 5	h. 3.5%

9. Describe the strength of gravity on each of the following planets (if any), where more black loops represent stronger magnetic fields. (κ=0.83)



Response Categories	N	Percent of Total
a. More magnetic fields = stronger gravity.	a. 69	a. 48.3%
b. Gravity and magnetism have no relationship.	b. 23	b. 16.1%
c. Gravity cannot exist without magnetic fields.	c. 11	c. 7.7%
d. Magnetic fields cancel out a planet's gravity (A is strongest).	d. 4	d. 2.8%

10. If you drilled a thin tunnel all the way to the center of the Earth, at which point would the gravity you feel be stronger? Explain.



Student responses to this question did not provide any new, previously undocumented information.

APPENDIX F

OPEN-RESPONSE FORM B WITH RESULTS

Questions about Gravity
Version B
Fall 2010

1. Why don't the planets fall into the Sun? (κ=0.94)

Response Categories	N	Percent of Total
a. Gravity	a. 33	a. 24.1%
b. Orbiting is like falling, centripetal forces.	b. 27	b. 19.7%
c. They are on an established orbit.	c. 13	c. 9.4%
d. Gravity from other bodies in the solar system is pulling them out.	d. 13	d. 9.4%
e. Sun's gravity isn't strong enough to pull them in, only to hold them in place.	e. 7	e. 5.1%
f. Rotation of planets or the Sun	f. 7	f. 5.1%
g. The planets have their own gravity that counteracts the Sun's.	g. 6	g. 4.4%
h. Sun's magnetic field pushes out.	h. 5	h. 3.6%

2. Besides gravity, are there any other forces that hold us to the Earth's surface? (κ=0.84)

Response Categories	N	Percent of Total
a. Nothing else.	a. 77	a. 56.2%
b. Atmospheric forces.	b. 17	b. 12.4%
c. Rotation and orbital dynamics.	c. 9	c. 6.6%
d. Friction.	d. 5	d. 3.6%
e. More than one of the above factors	e. 4	e. 2.9%
f. Yes, maybe.	f. 4	f. 2.9%
g. Magnetic forces.	g. 3	g. 2.2%
h. Just provided random Physics terms	h. 2	h. 1.5%

3. The strength of gravity on top of Mt. Everest is 0.25% less than the strength of gravity at sea level. Why is this? (κ=0.83)

Response Categories	N	Percent of Total
a. Further from core where gravity is stronger.	a. 39	a. 28.5%
b. Higher elevation, greater distance from center of mass.	b. 32	b. 23.4%
c. Higher in atmosphere, closer to space.	c. 20	c. 14.6%
d. Less air pressure, air is thinner.	d. 14	d. 10.2%

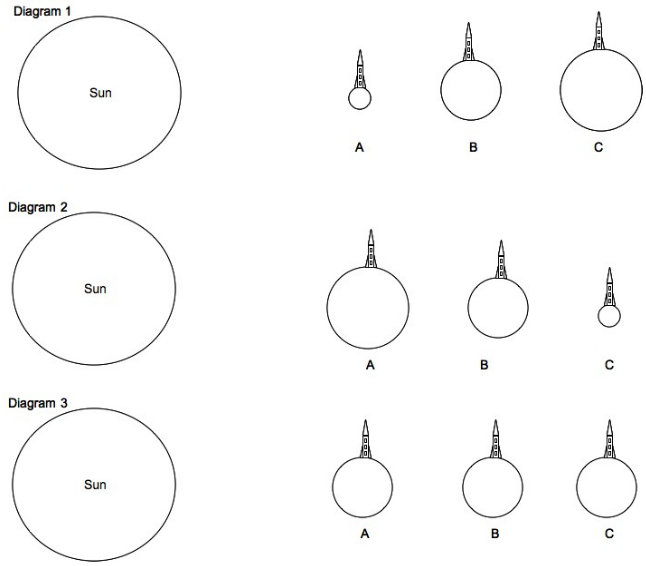
4. How would you compare the strength of gravity at the surface of Pluto with the strength of gravity at the surface of the Earth? Explain. (**κ**=0.91)

Response Categories	N	Percent of Total
a. Weaker, smaller and less mass.	a. 61	a. 44.5%
b. Weaker, further from Sun.	b. 26	b. 19.0%
c. Need more information, describes *how* to compare	c. 9	c. 6.6%
d. Weaker, no explanation	d. 7	d. 5.1%
e. Greater, for various reasons	e. 7	e. 5.1%
f. Weaker, Pluto has less atmosphere	f. 5	f. 3.6%
g. They're different.	g. 4	g. 3.0%
h. The same.	h. 2	h. 1.5%

5. When the Apollo spacecraft passed behind the Moon from our view on Earth, how did the gravity between the Earth and the spacecraft change (if at all)? Explain. (**κ**=0.83)

Response Categories	N	Percent of Total
a. No change, gravity can pull through objects.	a. 31	a. 22.6%
b. Moon's gravity takes over, not feeling Earth's gravity anymore.	b. 12	b. 8.8%
c. Moon blocked/stifled Earth's gravity.	c. 11	c. 8.0%
d. Now BOTH Earth and Moon are pulling.	d. 9	d. 6.6%
e. No gravity in space to begin with.	e. 5	e. 3.6%

6. For each of the diagrams below, assume all planets are made of the same material and decide which planet would be easiest for the rocket to "take off" from. Explain your choices.

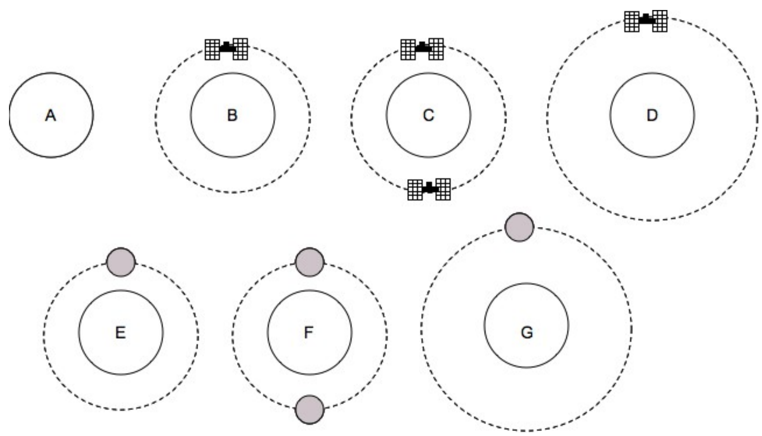


Student responses to this question did not provide any new, previously undocumented information.

7. The following planets are viewed from above, with more arrows representing a faster rotation. Describe the strength of gravity (if any) on each planet and how it compares to the other planets. (κ=0.83)

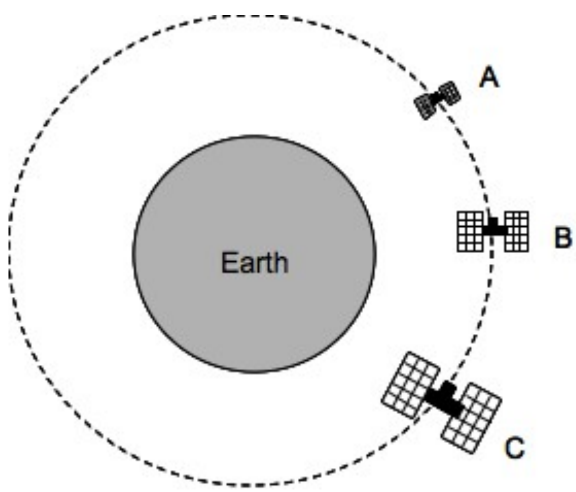
Response Categories	N	Percent of Total
a. More rotation = more gravity.	a. 48	a. 35.0%
b. All have the same gravity.	b. 29	b. 21.2%
c. Gravity is non-existent without rotation (A has no gravity).	c. 18	c. 13.1%
d. The direction of rotation matters for gravity.	d. 9	d. 6.6%
e. More rotation = less gravity.	e. 9	e. 6.6%

8. Compare the surface gravity on each of the following planets with different orbiting bodies. Planet A has nothing orbiting it; B-D have artificial satellites, and E-G have moons. Also, which planets would you expect to have tides?



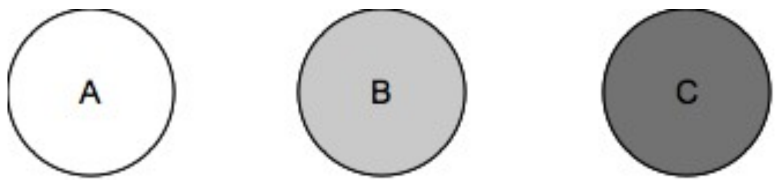
Response Categories	N	Percent of Total
a. Gravity must be stronger to keep heavy objects near.	a. 30	a. 21.9%
b. Gravity must be strong to hold objects that are far away.	b. 15	b. 10.9%
c. Same, surface gravity doesn't depend on what is orbiting.	c. 14	c. 10.2%

9. Compare the strength of the force of gravity for each satellite. Note that Satellite B is twice as big as Satellite A and Satellite C is three times as big as Satellite A. (**κ**=0.87)



Response Categories	N	Percent of Total
a. Larger, more massive satellites experience a stronger force of gravity.	a. 72	a. 52.6%
b. Same, no explanation.	b. 35	b. 25.5%
c. Smaller satellites experience a stronger pull of gravity (b/c more easily affected).	c. 11	c. 8.0%
d. Same, gravity pulls on all objects equally.	d. 3	d. 2.2%
e. Same, they are at the same distance.	e. 2	e. 1.5%

10. In this diagram, darker planets have a higher density. Which planet would have taller mountains? Why? (**κ**=0.86)



Response Categories	N	Percent of Total
a. A, lower density means more malleable.	a. 33	a. 24.1%
b. B, it's in between, just right	b. 28	b. 20.4%
c. C, more material to make mountains.	c. 22	c. 16.1%
d. A, has less gravity.	d. 15	d. 10.9%
e. C, more geologic activity	e. 7	e. 5.1%
f. Need more information.	f. 6	f. 4.4%

APPENDIX G

OPEN-RESPONSE FORM C WITH RESULTS

Questions about Gravity
Form C Questions and Results
Fall 2010

1. How far away from the Earth to you have to go to feel zero gravitational force from the Earth? (**κ**=0.95)

Response Categories	N	Percent of Total
a. Outside of the Earth.	a. 15	a. 46.9%
b. On the scale of much larger than the solar system.	b. 5	b. 15.6%
c. On the scale of nearby bodies and the solar system.	c. 4	c. 12.5%

2. What determines the strength of the gravitational interaction between two objects? (**κ**=0.95)

Response Categories	N	Percent of Total
a. Mass and/or distance, but incomplete.	a. 16	a. 50.0%
b. Complete: Mass of both objects and distance between them.	b. 11	b. 34.4%
c. Variables other than mass and distance.	c. 2	c. 6.3%

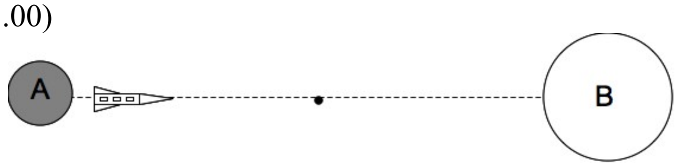
3. Why does the speed of a planet change over the course of its orbit around the Sun?

Student responses to this question did not provide any new, previously undocumented information.

4. How would the motion of a baseball thrown upward from the surface of the Moon differ from that on Earth? Describe how fast it will go and where it will end up. (**κ**=0.93)

Response Categories	N	Percent of Total
a. Will land back on surface.	a. 20	a. 62.5%
b. Will travel away from the Moon.	b. 4	b. 12.5%
c. Will float.	c. 4	c. 12.5%

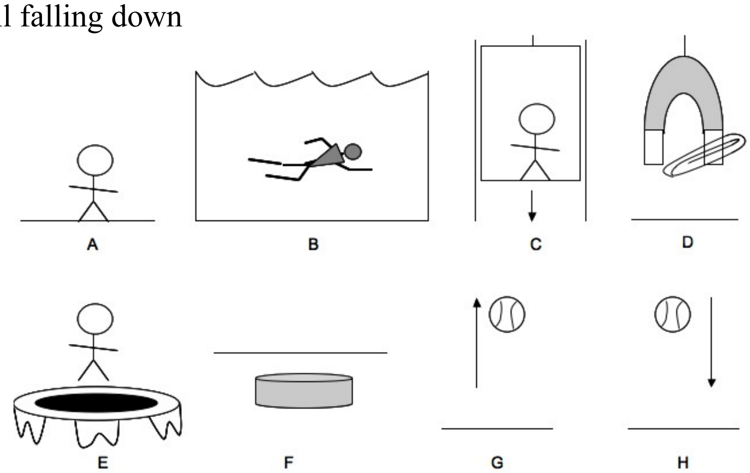
5. Assume that planets A and B have the same mass M, but planet A is made of a denser material and planet B is bigger. Make an "X" about where a rocket that is launched from planet A begins to experience a stronger force of gravity from planet B. The black dot shows the halfway point for reference. Explain your choice. (**κ**=1.00)



Response Categories	N	Percent of Total
a. Recognizes that pull of gravity would be equal at halfway point	a. 13	a. 40.6%
b. Closer to B, no explanation	b. 8	b. 25.0%
c. B has more gravity because it is larger.	c. 6	c. 18.8%
d. A has more gravity because it is more dense.	d. 3	d. 9.4%
e. Closer to A, no explanation	e. 1	e. 3.1%

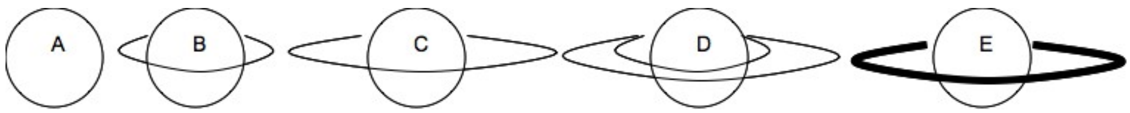
6. What is the direction (if any) of the gravitational force experienced by each object near the surface of Earth?

   A. person standing on the ground
   B. person swimming under water
   C. person on an elevator gaining speed and going down
   D. paper clip stuck on a magnet
   E. person jumping up on a trampoline
   F. brick buried under the ground
   G. ball thrown up
   H. ball falling down



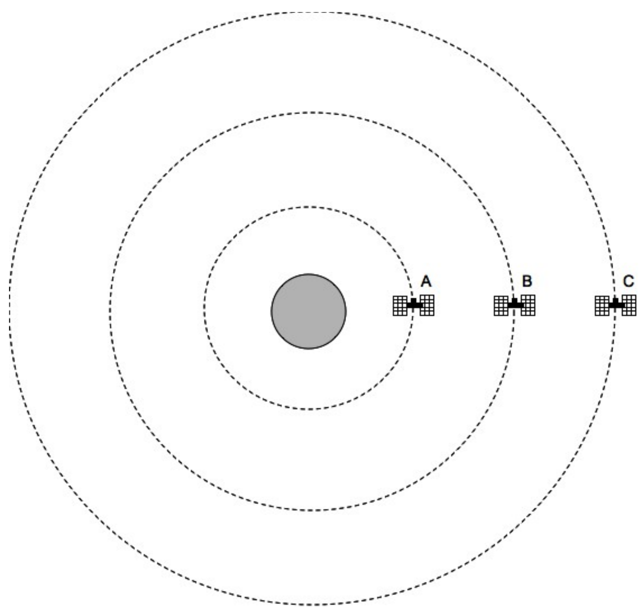
Student responses to this question did not provide any new, previously undocumented information.

7. Compare the force of gravity at the surface of each of the following planets that have different ring structures. (**κ**=1.00)



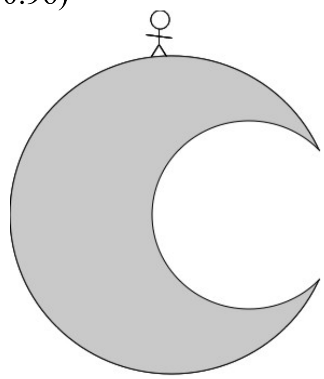
Response Categories	N	Percent of Total
a. More material around a planet = stronger gravity.	a. 13	a. 40.6%
b. Less material around a planet = stronger gravity.	b. 5	b. 15.6%
c. Planets with rings far away = more gravity.	c. 4	c. 12.5%
d. All the same.	d. 2	d. 6.25%

8. Satellite A is twice as close to the central planet as Satellite B and three times closer than Satellite C. Compare the gravitational force, if any experienced by each satellite with the planet.



Student responses to this question did not provide any new, previously undocumented information.

9. A person is living on a planet where there has been a huge hole carved out. Draw an arrow that shows the direction of the gravitational force he experiences. Explain your choice. (κ=0.96)



Response Categories	N	Percent of Total
a. Toward new center of mass.	a. 11	a. 34.4%
b. Toward geometric center.	b. 7	b. 21.9%
c. Toward left surface.	c. 4	c. 12.5%
d. Toward the gap.	d. 3	d. 9.4%

10. Two astronauts are floating in space far away from any planets or stars. Do they experience any gravitational force? If so, from what and will its effects be measurable? If not, why not? (κ=1.00)

Response Categories	N	Percent of Total
a. Yes, from each other.	a. 12	a. 37.5%
b. Yes, from something else.	b. 6	b. 18.8%
c. No, there is no source nearby.	c. 6	c. 18.8%
d. No, there is no gravity in space.	d. 3	d. 9.4%
e. No, no explanation	e. 4	e. 12.5%

APPENDIX H

STUDENT INTERVIEW NOTES FOR OPEN-RESPONSE QUESTIONNAIRES

Student Information	Interview Notes
F10-01 Form B Undeclared	- Planets don’t fall into the Sun because they’re in orbit. The angle they are falling is too steep, so they fall past it and circle around. Or it is because they’re too far away? - Gravity is the only force that holds us to Earth. - Mt. Everest has weaker gravity because it is farther from the center of Earth, therefore farther from the center of gravity. Gravity comes from mass, and that’s where the mass is centered. If traveling through a tunnel through center of Earth, one would float, like the inverse of Mt. Everest; i.e. weaker gravity because some mass would be above you. - Pluto has weaker gravity than Earth because it is less massive, similar to the Moon. - Earth still affects Apollo even when behind the Moon. Maybe the Moon added to gravity. - Rotation has nothing to do with gravity. - Stuff orbiting farther away indicates stronger gravity. One could look at planets in our solar system and how far out their moons are to determine their gravity. Mass of the moons and how fast they are orbiting may also matter. If the planet had more gravity, the moon could orbit faster. Closer orbits are always faster, like with elliptical orbits. Pluto is the slowest and Mercury is the fastest. - Orbiting artificial satellites of different mass would experience the same gravity, because they are not big enough to affect gravitational pull. If they were moons, though, then maybe the bigger one would experience more gravity. - Maybe easier for mountains to form on low-density planets because it is easier to push up mountains. But terrestrial planets are the only ones with mountains, so maybe it requires more density.
F10-02 Form B Undeclared	- The Sun’s gravity is strong enough to hold us in place but not pull us in. It’s not a set boundary. Objects must be very far away to start orbiting. - Originally thought that rotation also keeps us on Earth, but once being interviewed decides that spinning would throw us off, so decides that gravity is the only force keeping us on Earth. - Less gravity at Mt. Everest because it is far from center of Earth, where there is the most dense mass and gravitational pull.

	- Gravity is related to magnetism. Everything is attracted to each other gravitationally. Maybe magnetically, too. - Not sure if mass creates gravity, but larger = more gravity. Some objects don't have a lot of gravity; it depends on how dense they are and the composition. - Gravity between Apollo and Earth is the same, but now the Moon is pulling too, like the tides. The pull gets stronger as the spacecraft gets closer to the Moon. - Not sure if rotation affects gravity. Spinning throws objects off, but if the planet is held together with gravity, there must be more gravity to prevent things from falling off. - Could looks at planets in the solar system and the objects that orbit it, including their masses and how far away they are, and could deduce the planet's gravity. - More massive satellites experience stronger gravity. Can't remember if it's the mass or the distance that's squared. - Tries to think of a relationship between density and mountain formation. Maybe denser objects with more gravity have more geologic activity.
F10-03 Form A Undeclared	- There is no gravity in space. It gradually decreases from Earth's surface. - The Moon has its own gravity. It doesn't feel Earth's gravity. The space station isn't a big enough mass to have its own gravity. - Force of gravity is greater at the center of a mass. That part doesn't spin as fast. - No relationship between gravity and atmosphere. Gravity has to do with mass and interaction with the next closest mass. - Everything has a gravitational field. Heavier objects have more gravity than lighter objects. - No change in gravity during an eclipse because the Moon and Earth are more or less the same distance apart. The position of the Sun doesn't matter. - People at Earth's poles experience a stronger force of gravity, because Earth is squished at the poles, so they are closer to the core. There is also more gravity for moons orbiting near the poles of a planet. - Denser planets have stronger gravitational pulls. - Magnetic fields don't change the strength of gravity on a planet's surface, but they change the extent of gravity because they push out, weakening gravity.
F10-04 Form A	- Astronauts float because they are too far from a massive body; there is weak gravity. They have broken out of the

History		atmosphere, where Earth's gravity is strong enough to hold. Atmosphere is an indicator of where gravity works.
	-	Pressure is an indicator of how much gravity there is. Traveling down a tunnel, pressure would increase. Gravity depends on distance too, so it seems that gravity would be stronger.
	-	No cause/effect relationship between gravity and atmosphere. Big atmosphere indicates the planet must have strong gravity.
	-	Objects much be pretty heavy to create their own gravitational field, but it's more to do with relative mass and density. If the Sun was larger, like a red giant, it would push objects out.
	-	A person standing at the equator experiences a marginally stronger gravitational force compared to at other latitudes because Earth bulges around the equator. It's bulkier around the equator; there is more material. There would be no difference if Earth were a perfect sphere.
	-	Gravity is calculated by comparing mass and distance ratios.
	-	Stronger magnetic fields indicate that gravity is more intense. Maybe the core does have some special purpose.
	-	Traveling down a tunnel, gravity would increase because you would be surrounded by denser matter.
F10-05 Form B Undeclared	-	The planets *are* technically falling into the Sun. That's what orbiting is. It's slow, so there's no need to worry, but they will eventually fall into the Sun.
	-	Both Earth and the Sun's gravity hold people to Earth.
	-	The center of the planet is denser. The core is the source of gravity. Earth's gravity (not Sun's) is less on a mountaintop.
	-	Pluto is smaller than Earth so it has weaker gravity.
	-	Apollo passing behind the Moon didn't change the strength of gravity it experiences from Earth.
	-	Rotation speed is not enough information to determine strength of gravity, but it is probably related to gravity. The speed probably wouldn't be random.
	-	Strength of Earth's gravity would be the same on satellites of different mass.
	-	Density of a planet is not enough information to determine how mountains form.
F10-06 Form A Architecture	-	No gravity in space. Maybe Earth's gravity ends past the Moon, since it's orbiting. Or, maybe gravity is everywhere. Maybe something must be orbiting to experience gravity.

	Mars only feels gravity from the Sun. - Strength of gravity wouldn't change in a tunnel. Earth's gravity originates from Earth as a whole, so it's the same everywhere on Earth. Starts to die off once you get to space. - Maybe the edge of the atmosphere is at the same location as the boundary of gravity by coincidence. - An object experiences gravity in relation to the masses that surround it. In space, keys would float toward a person. Apollo mission still feels attraction to Earth because Earth is more massive than the Moon. The Moon's position is determined by Earth's gravity. - More magnetic fields, stronger gravity. Maybe the boundary is a boundary of magnetic fields. - Comments that he mostly just knows of gravity in terms of things falling down; it's difficult to make the jump to how things in space are affected.
F10-07 Form B Health Promotion	- Doesn't know why planets don't fall into the Sun. We're secure in our space based on mass and the pull of the Sun. Maybe Earth isn't massive enough to fall in. Saturn is farther away, so it has more mass to not fall in. - Mass creates gravity. Maybe other forces hold us to Earth's surface, but doesn't know what. - Mt. Everest is farther from the core, where energy comes from. Earth is denser at the center, so gravity might be stronger at the center. How tightly packed the material is matters, but not sure. - Pluto is a Jovian planet so it is less dense. Earth is terrestrial. - Knows that Earth's pull extends up past the atmosphere to a certain extent. If there is no atmosphere then gravity would stop inside because it needs a substance to travel through. Guessing, thinks there must be a contradiction. - Doesn't know if there is a relationship between distance from the Sun and gravity. - Not sure if rotation affects gravity. Maybe it does since the question is being asked in the first place. - Gravity would hold back mountains from being formed.
F10-08 Form B Architecture	- Planets don't fall into the Sun because of the centripetal acceleration; the same reason you fly to the outside of a turn. - Gravity is caused by the mass of Earth. Atmosphere pushes down, but the push is based on gravity in the first place. - Gravity is weaker on top of Mt. Everest because of a combination of being farther away from the center of mass

	and being on the outside of a turn. - Pluto has less gravity because it has less mass. - Earth would still pull on Apollo even when it was behind the Moon, but the Moon would also pull. - Most gravity on non-rotating planets. But, if the planets are all the same size and same mass, then gravity should be the same. - More massive satellites would have more effect on gravity, so they would have to be in a faster orbit. Wait, no, less massive satellites would have to travel faster to maintain the same orbit. - Less dense means less gravity, allowing mountains to grow taller because there is less resistance.
F10-09 Form B University Studies	- Planets don't fall into the Sun because their magnetism counteracts the Sun's magnetism. Gravity is moderated by magnetism. Centrifugal force wouldn't hold them close, because that throws them out. - We are held to Earth probably by some complicated equation that balances many forces. Must account for Sun and Moon. - Sun is pulling harder on the top of Mt. Everest because a person would be closer to the Sun there. Less Gravity because Earth's core is farther away; there is a thinner atmosphere, indicating weaker gravity, even though this doesn't necessarily determine gravity. References Venus' atmosphere. - A planet's distance from the Sun determines it's gravity. Mars is less than Earth, Venus is more. Maybe Jupiter is more because it is gaseous and has more mass. It could be described by a complicated equation. - Maybe gravity between Apollo and Earth was a little weaker because the Moon was in the way. - More rotation means less gravity, but you can't determine gravity from just rotation alone. - There is no gravity in space, just magnetic fields that pull.
F10-10 Form A Business Marketing	- Astronauts float because there is no gravity in space. "Space" is when they pass the magnetic pole, actually when they leave the atmosphere. But, the Moon orbits, so it must feel the same gravity that astronauts feel. Don't know about past the Moon. - Traveling down a tunnel would not change gravity. You would just get hotter. Earth has gravity because of its mass and its rotation. You would still be feeling all of the mass. Gravity is the same for planes flying in the sky, too,

	because they are still in the atmosphere. It would weird if it wasn't. Nothing feels lighter when you're in a plane. - Gravity is greater if a planet has an atmosphere. Atmosphere creates more gravity. - Objects must be very heavy to create their own gravitational field. Earth is about the minimum; don't think the Moon has gravity because not orbits it. Mercury and Venus are smaller and they don't have moons. Maybe they have gravity, but not a field. To have a field means to be affecting other objects' motion. - Gravity doesn't change during eclipse. If it does, it's very little, because maybe the alignment of the Sun and Moon tug on Earth a little more. - Gravity is 9.8m/s^2 everyone on Earth. - Earth's gravity probably doesn't depend on the position of the Moon, because if it did, then it would show up in physics problems. - Denser planets cause more gravity. A rocket close to the surface of a large, low-density planet would not be affected by gravity. - Stronger magnetic fields and stronger gravity are intertwined. You need to know both mass and the strength of a magnetic field. - No one knows how strength of gravity would change in a tunnel to the center of Earth, because no one has ever been there. Maybe it would be the same.
F10-11 Form A University Studies	- Astronauts float because they could be far enough away to not be affected by Earth's gravity. Earth's gravity is confined to just our planet. When you leave you hit a certain point where it stops. Although, in order to orbit, there must be gravity. Maybe things don't fall because of our atmosphere. - Earth's center has the highest concentration of gravity. - A thicker atmosphere offsets gravity. It blocks gravity from other bodies in the solar system, but if gravity comes from our own planet, then that's not the case. Not sure. - All objects have some gravity. Mass and distance defines the amount. This is just common sense. It doesn't make sense for it to apply to one thing but not another. - Lunar eclipse doesn't change gravity. Light doesn't have to do with gravity. - Person at the equator experiences strongest force because they are at the fastest spinning part and closest to the Moon, so they have the most things pulling on them. If gravity

	comes from Earth itself, then they would all be the same, but under the impression that gravity comes from outside forces. - A rocket orbiting around a larger planet experiences larger force because it is closer, and gravity depends on mass and distance. Although, if it's the distance from the center, then the force would be the same. - Magnetic fields counteract gravity. Gravity pulls in, magnetic fields push out.
F10-12 Form B Undeclared / University Studies	- Planets don't fall into the Sun because an orbit has been established, like a toilet or coin funnel – there is an outward force that matches gravity. - Friction prevents us from leaving Earth; it's why we need so much rocket fuel, there is so much air resistance. Maybe air holds us to Earth, like pressure. - The farther from an object, the less gravitational attraction. - Pluto's gravity is weaker than Earth's because it is smaller. - When Apollo is near the Moon the strength the Moon is more pronounced, but there is still a force with Earth. - More rotation means less gravity. Analogy to a merry-go-round: it's harder to hold on to. - Objects with larger mass have more attraction, i.e. it's more of a tug of war. It would be more difficult to keep a more massive satellite aloft; it would need a higher orbit. The distance might be a law of squares.
F10-13 Form B Film	- Planets don't fall into the Sun because of inertia. They go in a straight line and gravity pulls them in so they are constantly falling in orbit. If they have enough speed, they won't ever fall into the Sun. - Gravity holds us to Earth because there is no separation between Earth and us. Astronauts don't feel gravity after a certain distance. Past the atmosphere there is very little gravity. - The core of a person on top of Mt. Everest is farther from the core of Earth. The entire thing makes gravity, but the center is the point where gravity pulls. - Mass determines gravity, so Earth has more gravity than Pluto. - Apollo was still puled by Earth; it doesn't matter that the Moon is blocking it. - Gravity only depends on mass and distance. Rotation doesn't matter. Gravity is a separate thing. - Larger satellites experience a larger force. They need more force to pull them in.

	- Forming mountains is easier when there is less gravity.
F10-14 Form C Psychology	- To feel zero gravitational force from Earth, one must go right outside the atmosphere, to space. Planets have their own pull. The Sun's gravity works with the whole solar system. Mars and Earth to not feel each other's gravity. Planets just feel the Sun's gravity. - Moon doesn't experience Earth's gravity. It has it's own. Orbits because the two forces work together. Don't know why the International Space Station orbits. - Mass and distance determine the strength of the gravitational interaction between two objects. - Speed of planets changes because they orbit in ellipses. Distance and gravitational pull change. - A baseball thrown upward from the Moon would float, stall out a little, and then come back down slowly. It would go higher than on Earth. - Higher density means more gravity, but this effect is not a big factor. - At first thought that gravity was in the direction of motion of objects near Earth's surface, but after questioning decides that it's always down. - Planets with rings farther out have more gravitational pull. Earth has a pull that's only so far, but if rings were added it would expand our influence. Rings are part of a planet; they make the planet bigger. It doesn't matter how many rings there are, just how dense/thick they are. - Satellites closer to Earth feel a stronger gravitational force, and they will orbit faster. - A planet with a hole carved out will pull a person toward the hole to try to seal itself up. The hole has more gravity than the rest of the planet. - Initially thought that two astronauts in space would feel a force toward each other because they both have mass, but decided that they are not massive enough so there wouldn't be a gravitational pull. They would need to be at least the size of the Moon.
F10-15 Form B Geology	- Planets don't fall into the Sun because the universe is expanding. Although, this can't be true because galaxies stay their same shape. Something else must counteract gravity, but not sure what. - Weight of the atmosphere also holds us to Earth's surface. If the atmosphere weren't here, we would feel less force. Gravity is not determined by atmosphere, just mass. - The top of Mt. Everest is farther away from the center of

	Earth, i.e. the densest point. The center is just where the average of the mass is. If Earth were lopsided it would pull in a different direction. - Earth is larger than Pluto so it has more gravity. - Gravity doesn't depend on rotation. Rotation gives another force. Planets that rotate faster want to throw us off. - More massive satellites would have more gravity. Not sure if gravity scales with mass, i.e. 3x the mass implies 3x the gravity. - More force to build up mountains against strong gravity. But planets need to be dense enough to hold the weight up.

APPENDIX I

CONCEPT INVENTORY VERSION 1

Gravity Concept Inventory
Version 1

Please answer the following questions about gravity to the best of your ability. This questionnaire is for research purposes only, so your performance will NOT affect your grade in any way. Your student ID number will be used for matching your responses to other data, but your information will be destroyed after the study is over.

1. Besides gravity, what other factors, if any, hold us to the surface of the Earth?
   a. Magnetism from the Earth's magnetic field
   b. Air pressure from the Earth's thick atmosphere
   c. Rotation as the Earth spins and rotates around the Sun
   d. No other factors; gravity is the only one.

2. Gravity is primarily responsible for which one of the following?
   a. Spherical shape of the Earth
   b. Number of craters on the Moon
   c. Brightness of light created by a star
   d. Presence of magnetic fields on the Sun

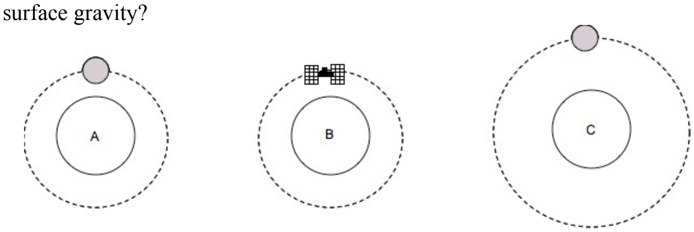
3. How would the gravitational force between the Earth and Moon change if the Moon moved out to *ten times* the distance of its present location?
   a. Would be unchanged
   b. Would decrease only slightly
   c. Would be half as strong
   d. Would decrease a huge amount.

4. Two astronauts are floating in space far away from any planets or stars. What is the direction of the gravitational force that they experience, if any?

   a. Toward each other, because they are being pulled by each other's gravity.
   b. Away from each other because they are pulled by planets and stars in the distance.
   c. They experience a gravitational force, but it cannot be determined in what direction.
   d. They do not experience a gravitational force because there is no source nearby.

5. Astronauts appear weightless in their spacecraft because…
   a. They are falling at the same rate as the spacecraft.
   b. There is very little gravity in space / outside earth's atmosphere.
   c. They are too far away from Earth or any massive body.
   d. There is no gravity inside the spacecraft.

6. Planets A, B and C have the same mass, but A and C each has a heavy moon orbiting it, and B has a lightweight satellite orbiting it. Which planet has stronger surface gravity?

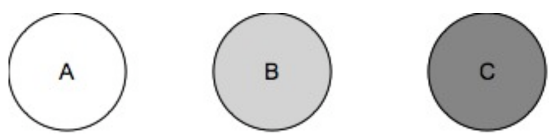


   a. Planet A, because it is able to hold a heavy object close to it.
   b. Planet B, because only lightweight objects can orbit it.
   c. Planet C, because it is able to hold a heavy object that is far away.
   d. All have the same gravity, because they are all the same mass.
   e. Cannot determine.

7. What causes/creates gravity?
   a. Any material with mass
   b. Solid material
   c. Dense material
   d. Air pressure

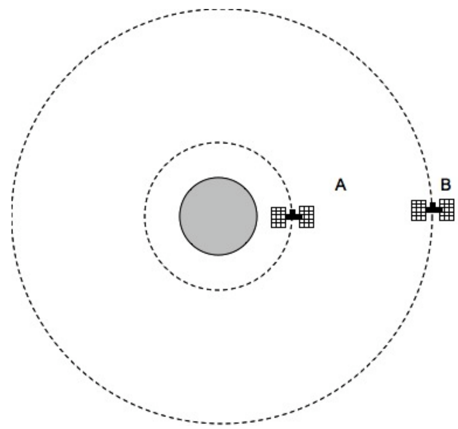
8. Why doesn't the Sun's gravitational pull cause the planets to fall into the Sun?
   a. It is only strong enough to hold them in orbit.
   b. The planets' own gravity is pulling back.
   c. The Sun's magnetic fields are pushing out.
   d. The speed of the planets flings them out.
   e. The Sun's rotation pushes the planets out.

9. For the following planets, a darker shade indicates a higher density. Which planet would have taller mountains?



   a. Planet A, because it has less mass, so less force pulling down.
   b. Planet B, because it is neither too dense nor too light.
   c. Planet C, because it has more material to form mountains.
   d. Same size mountains on all planets, because they are the same size.
   e. Cannot determine.

10. Satellites A and B are identical, except A is orbiting the planet *three* times closer than Satellite B. Compare the force of gravity on each satellite by the planet.



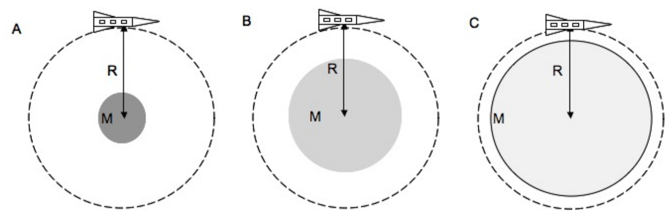
   a. A experiences a stronger force than B, but can't tell how much stronger.
   b. A experiences less than 3 times the force that B experiences.
   c. A experiences exactly 3 times the force that B experiences.
   d. A experiences more than 3 times the force that B experiences.
   e. A and B experience the same force.

11. Why does the Earth have gravity?
   a. It has mass.
   b. It has a magnetic field.
   c. It is so large.
   d. It is so dense.
   e. It rotates.
   f. It has an atmosphere.

12. Suppose the Sun shrank in size but its mass remained the same. What would happen to the orbit of the Earth?
   a. Earth's orbit would shrink.
   b. Earth's orbit would expand.
   c. Earth would fly off and leave the solar system.
   d. Earth's orbit would be unaffected.

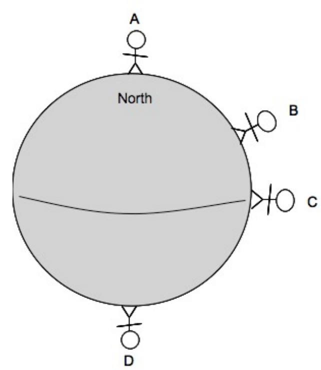
13. Is there a relationship between a planet's gravity and its atmosphere? If so, what is the relationship?
   a. Yes, the atmosphere represents the boundary of gravity.
   b. Yes, a thicker atmosphere causes stronger gravity.
   c. Yes, gravity holds the atmosphere to the planet.
   d. No relationship.

14. Each of the identical rockets below is orbiting at a distance R from a host planet of mass M. The darker the planet, the denser the material. Which rocket experiences a stronger force of gravity?



a. A, because higher density means stronger gravity when distance to center of mass is the same.
b. B, because it is of medium density and the rocket is somewhat close to the surface but not touching it.
c. C, because smaller distance to the surface of the planet means stronger gravity when mass is the same.
d. All the same, the stronger gravity from denser planets is compensated by the rocket's smaller distance to the surface of the planet.
e. All the same, since each planet has the same mass and the rocket is at the same distance from the center of mass.

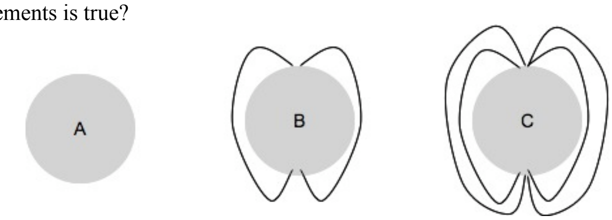
15. Assuming the Earth is perfectly spherical, which person standing on the surface of Earth experiences a stronger force of gravity? Why?



a. Both A and D, because they are near a stronger magnetic field
b. B, because he is affected by both the magnetic field and the rotation.
c. C, because he is at the fastest spinning part of Earth
d. D, because he is at the bottom of Earth and could fall off
e. All the same, because they are equidistant from the center of Earth.

16. How would you compare the strength of gravity at the surface of Pluto with the strength of gravity at the surface of the Earth?
    a. Weaker, it is further from the Sun.
    b. Weaker, it has less mass.
    c. Same, they both orbit the Sun.
    d. Stronger, its surface is closer to its core.
    e. Cannot determine.

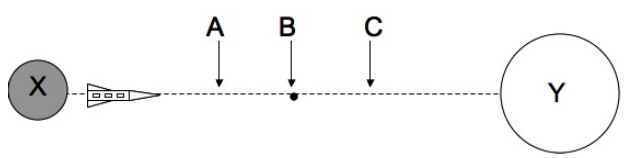
17. On each of the following planets, more black loops represent stronger magnetic fields, but the planets are otherwise identical. Which one of the following statements is true?



    a. Gravity cannot exist without magnetic fields.
    b. Magnetic fields increase a planet's gravity.
    c. Magnetic fields cancel out a planet's gravity.
    d. Gravity and magnetism have no relationship.

18. When the Apollo spacecraft passed behind the moon from our view on Earth, how did the gravity *between the Earth and the spacecraft* change (if at all)?
    a. Decreases, the spacecraft is not in view of Earth.
    b. Decreases, the moon's gravity stifles the Earth's gravity.
    c. Remains the same, Earth's gravity can pull through the moon.
    d. Remains the same, there is no gravity in space to begin with.
    e. Increases, the moon's gravity amplifies the Earth's gravity.

19. Assume that planets X and Y have the same mass M, but planet X is made of a denser material and planet Y is bigger. Where would a rocket that is launched from planet X begin to experience a stronger force of gravity from planet Y? The black dot shows the halfway point for reference.

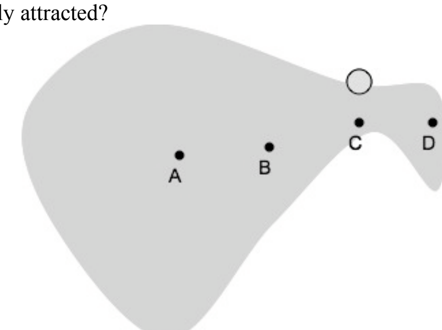


    a. At point A, because Y's gravity is stronger since it is bigger.
    b. At point B, because X and Y have the same mass, so the same gravity.
    c. At point C, because X's gravity is stronger since it is more dense.
    d. The rocket experiences zero gravity at each point, because it is in space.

20. An astronaut standing on the surface of the Moon has a pen in his hand and lets go. What happens to the pen?
    a. It falls to the surface of the Moon, but slower than on Earth because the Moon's gravity is weaker.
    b. It falls to the surface of the Moon, but faster than on Earth because the Moon has no atmosphere.
    c. It doesn't fall, but floats where it is because it isn't heavy enough to fall in the moon's weak gravity.
    d. It doesn't fall, but slowly drifts away because it is not gravitationally attracted to the Moon.

21. Where does the Earth's gravity originate?
    a. From the center point
    b. From the iron core
    c. From the surface
    d. From all points in Earth

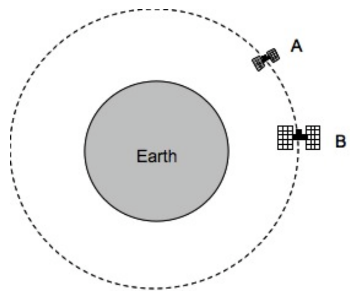
22. Consider a small, lumpy asteroid. Toward which point will the ball be gravitationally attracted?



    a. Point A, because it is the center of mass of the asteroid.
    b. Point B, because it is the geometric center of the asteroid.
    c. Point C, because it is directly below the ball.
    d. Point D, because it is at the edge nearest the ball.

23. How far away from the Earth do you have to go to feel *zero* gravitational force from the Earth?
    a. Outside the atmosphere
    b. The orbit of the Moon
    c. The orbit of Mars
    d. Outside the solar system
    e. An infinite distance

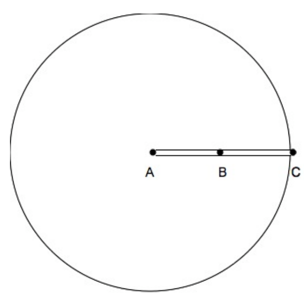
24. Satellite B is THREE times bigger than Satellite A. Which satellite experiences a stronger force of gravity with the Earth?



a. A experiences a stronger force than B, since it is easier to move.
b. B experiences a stronger force than A, since it has more mass.
c. A and B experience the same force, since they are at the same distance.
d. A and B experience no force, since they are in space.
e. Cannot determine.

25. Which one of the following is a way to change the strength of gravity you experience on the surface of the Earth?
a. Go swimming and float on the water
b. Ride an elevator down as it speeds up
c. Jump up from a trampoline
d. Climb a tall mountain
e. There is no way to change it.

26. Imagine a thin tunnel that reached to the center of the Earth. At which point (if any) would the gravity between you and Earth be strongest?



a. Point A, because the gravity from the outer layers is pushing down on you.
b. Point B, because you are being pulled by the gravity of the material at the center and pushed by the gravity of the material of the outer layers.
c. Point C, because gravity from the whole planet is pulling you.
d. The same for all points, because you have not left the planet.

27. Approximately what minimum mass must an object have to gravitationally interact with another massive object?
    a. Any mass will do
    b. About that of a person
    c. About that of a moon
    d. About that of a planet

28. The reason why the surface gravity on Saturn is similar to that on Earth is that:
    a. Saturn has the same mass as the Earth.
    b. Saturn has the same radius as the Earth
    c. Although Saturn is much more massive, objects on the surface are farther from its center
    d. All planets in the solar system have the same surface gravity

29. If thrown upward with the same force, how would the motion of a baseball on the moon differ from that on Earth?
    a. It would move slower, go up higher, then fall down.
    b. It would move faster, go up higher, then fall down.
    c. It would move slower and then float above the surface.
    d. It would move faster and leave the moon.

30. The strength of gravity on top of Mt. Everest is 0.25% less than the strength of gravity at sea level. Why is this?
    a. It is further from the Earth's center point.
    b. It is further from most of the mass of Earth.
    c. It is closer to space.
    d. The air is thinner there.

31. Which one of the following places is a zero-gravity environment?
    a. The Moon, because it has no atmosphere.
    b. The space shuttle, because it is outside of the Earth's atmosphere.
    c. An airplane at a constant height, because it is not falling.
    d. Just beyond the orbit of Pluto, because it is the edge of the solar system.
    e. Such a place does not exist.

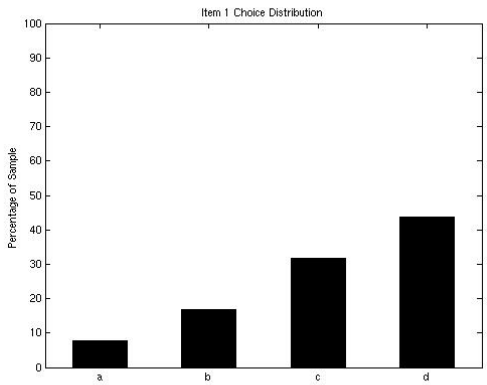
Demographic questions

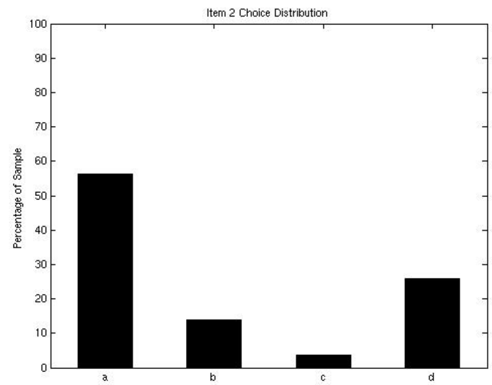
32. How many astronomy and/or physics courses have you taken before this one?
    a. 0
    b. 1
    c. 2
    d. more than 2

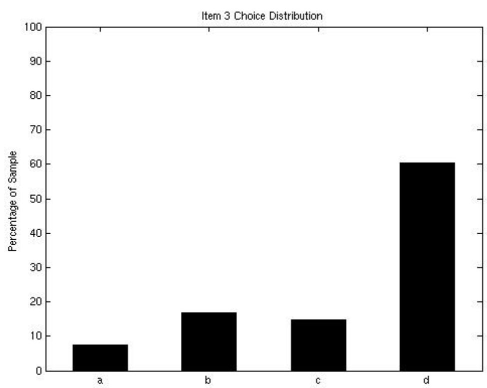
33. What is your gender?
    a. Male
    b. Female

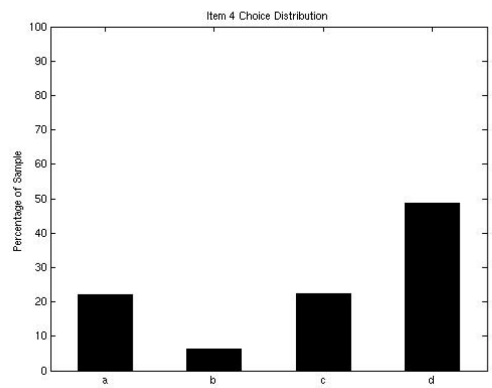
APPENDIX J

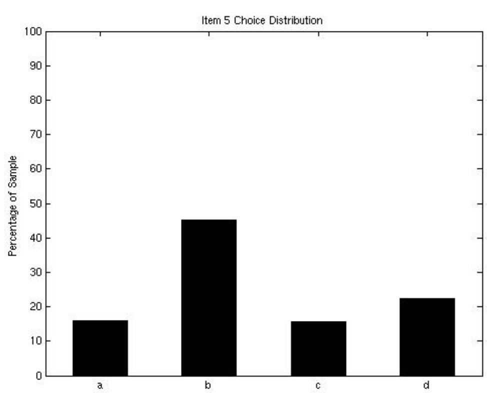
CONCEPT INVENTORY VERSION 1 STUDENT RESPONSES

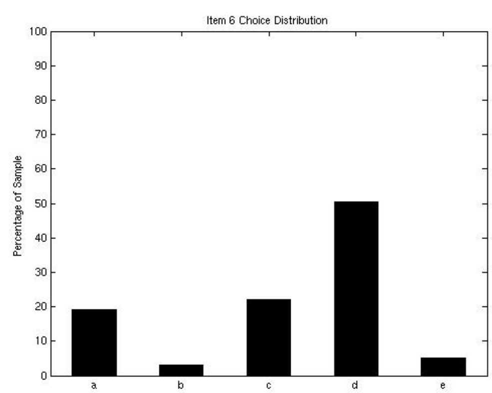

Item 1 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

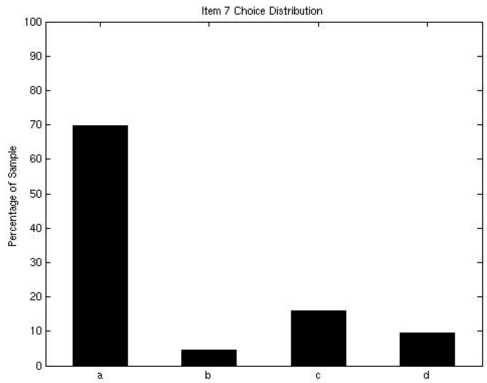

Item 2 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

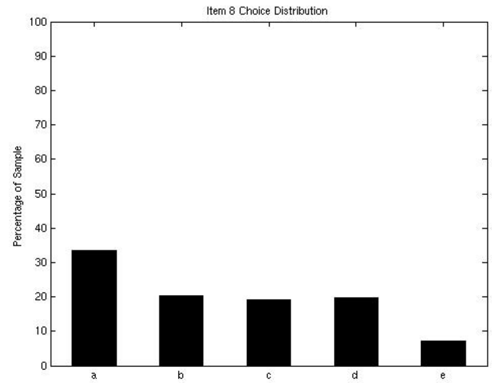

Item 3 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

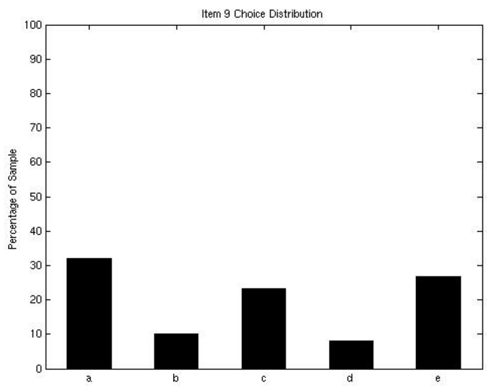

Item 4 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

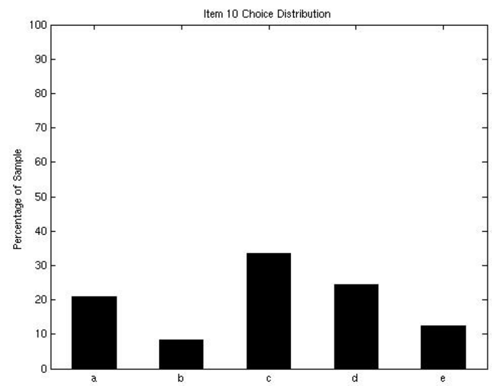

Item 5 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

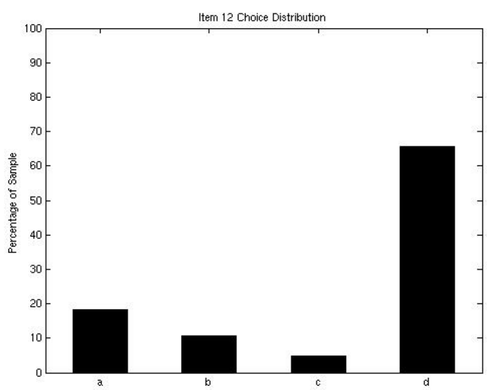

Item 6 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e

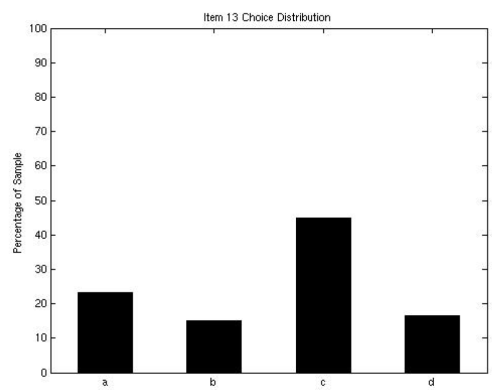

Item 7 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d

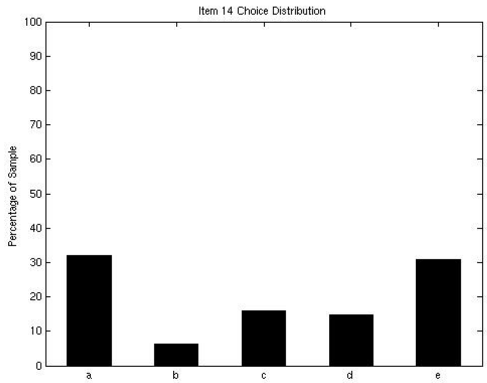

Item 8 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

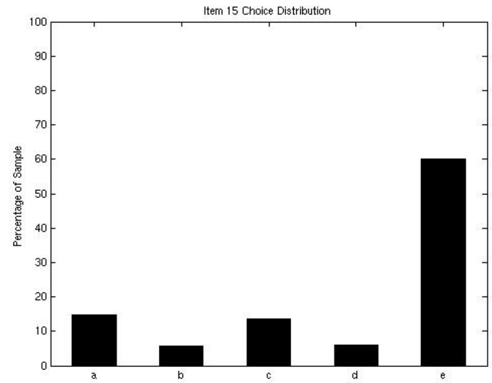

Item 9 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

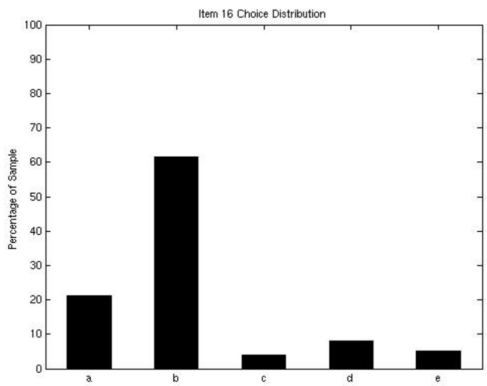

Item 10 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

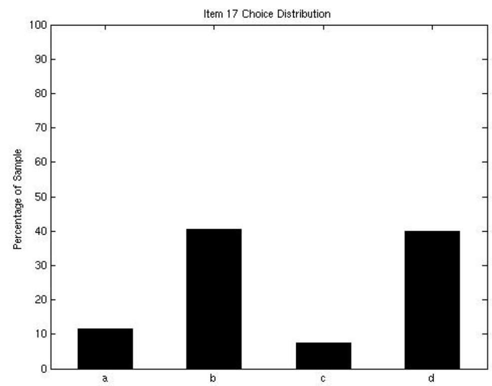

Item 12 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d

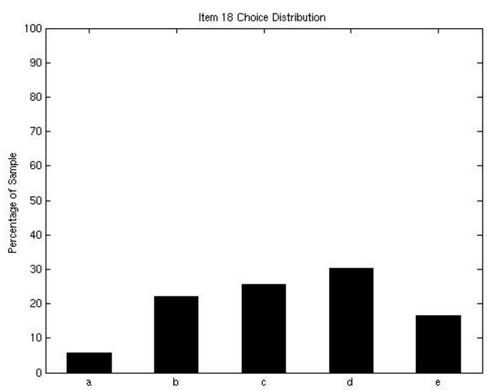

Item 13 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d

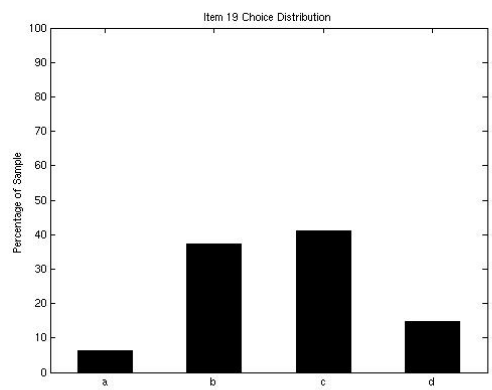

Item 14 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

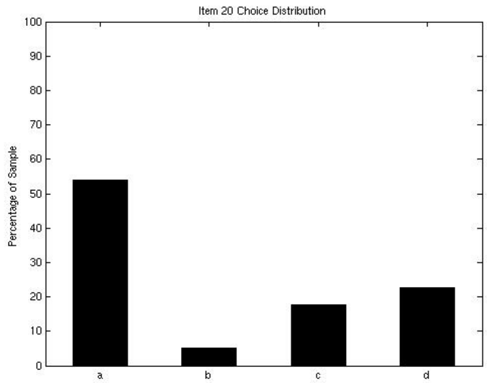

Item 15 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

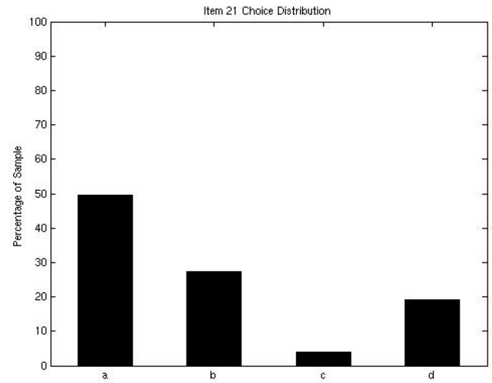

Item 16 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

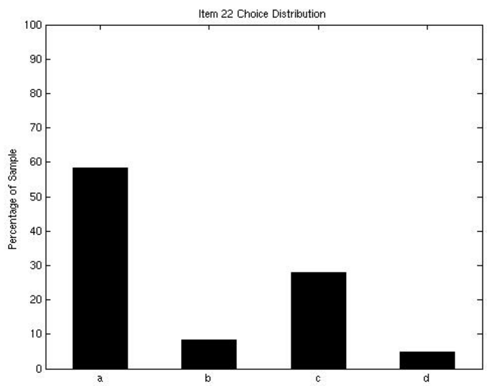

Item 17 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

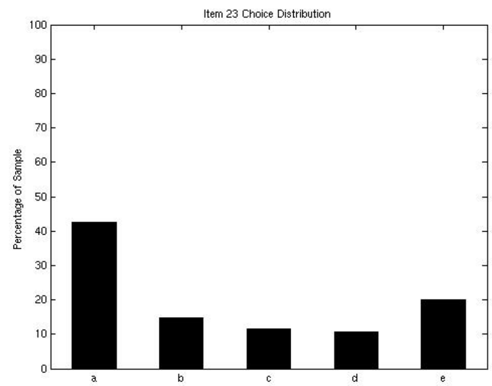

Item 18 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

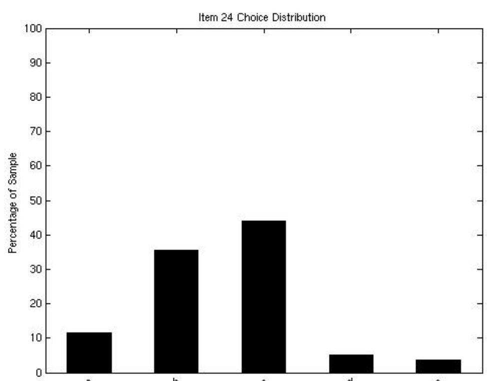

Item 19 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

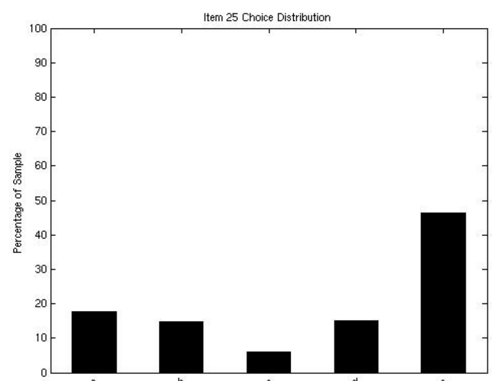

Item 20 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

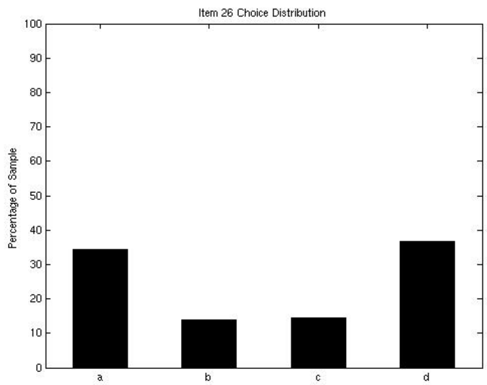

Item 21 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

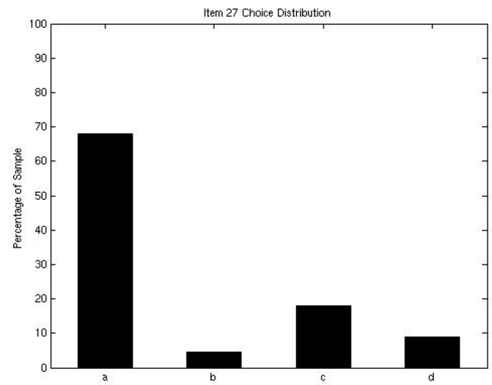

Item 22 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

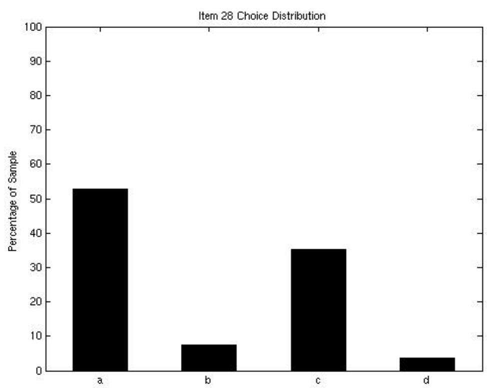

Item 23 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

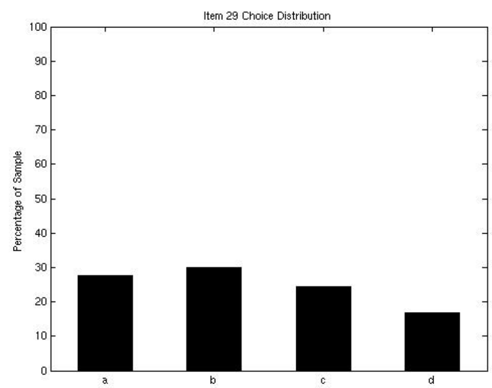

Item 24 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

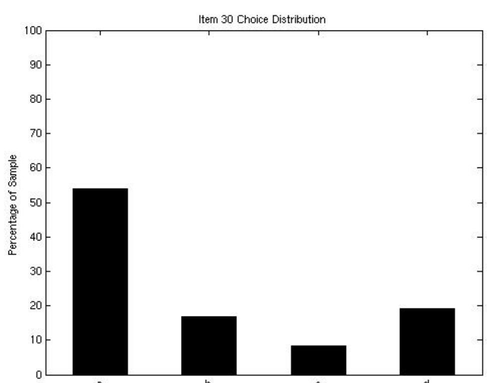

Item 25 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

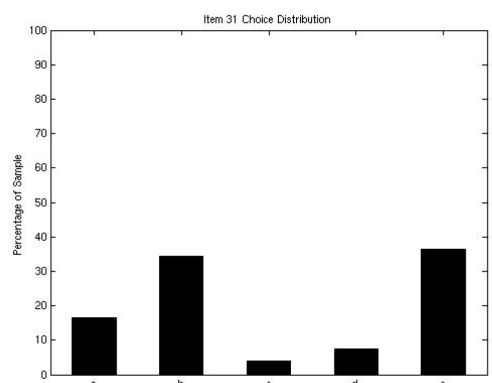

Item 26 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 27 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 28 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 29 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 30 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 31 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

APPENDIX K

MODIFICATION FROM VERSION 1 TO VERSION 2

Item	Modifications
1	Change "Rotation as the Earth spins and rotates around the Sun" to "Forces from the Earth's spinning motion" to make the choices closer to the same length.
2	Eliminate. There are too many nuances to this question, and it could be argued that there are multiple correct answers.
3	Change the distance from "ten times" to "twice" as far. The choices in Version 1 are too vague and open to interpretation.
4	Change wording so choices are similar in length.
5	Include that the spaceship is "Earth-orbiting" to more clearly target the idea of falling so choice 'd' cannot be argued to be correct.
6	Be more specific about mass. Do not use words like "heavy" and "lightweight," since these terms are not precise or useful in determining gravitational force from Newton's equation. Change "surface gravity" to "gravitational interaction."
7	Eliminate. This question implies that gravity is "stuff" rather than an interaction.
8	Eliminate. This question requires an understanding of angular momentum in addition to gravity.
9	Eliminate. Many geologic factors could affect the size of mountains. This question does not target a particular objective.
10	Refer to the force "between the satellite and planet" rather than "on the satellite by the planet." It is better to avoid any sort of indication of testing Newton's third law.
11	Ontological inconsistency: Earth cannot *have* gravity. Reword the stem to imply an interaction, i.e. exerting a force on objects on its surface.
12	Eliminate the choice "It is so large" to have the amount of choices that would fit on a standard bubble sheet. Also, this choice was somewhat confusing, i.e. does larger mean more massive?
13	Specify that the Sun would be denser to elicit the misconception about gravitational force being affected by density. Rather than refer to the size of Earth's orbit, ask directly about how the gravitational force would be affected. This more clearly targets the survey objectives.
14	Change distance labels from "R" to "D" to be more in line with what novice students would intuit. Eliminate case B. The "in between" case does not add any insight as to what students are thinking. Eliminate the reasoning in the choices.
15	Change "spherical" to "round" to reduce jargon. Eliminate reasoning in the choices.
16	Eliminate. It does not target a specific test objective.
17	Remove the diagram, as it is unnecessary for discussing magnetism. Re-word the question to more directly ask about the relationship between gravity and magnetism.

18	Eliminate and replace with a question that more effectively probes the misconception that the gravitational force can be blocked or stifled by an intervening object. Student interviews and responses indicated that students were focusing too much on the change in distance as a spacecraft orbits the Moon, which was unintentional.
19	Change the wording from the rocket beginning to experience a stronger force to just explicitly an equal force. Explicitly state that the black dot is the midpoint.
20	Change “lets go” to “releases it.” Remove reasoning in choices. Re-word choices to reduce reading load.
21	Eliminate. Again implies gravity is “stuff” that is created in a particular place.
22	Remove labeled points and replace with directional arrows
23	It could be argued that the gravitational force from Earth is *essentially* zero outside the solar system. To be more precise, modify the stem with the word, “Theoretically.”
24	Some students might choose “Cannot determine” when they simply do not know the answer. Replace this choice with “None of the above is correct” to force students to commit to a definite interpretation. Re-word choices to reduce reading load. Italicize, rather than capitalize, words/phrases that should be emphasized.
25	Eliminate. This question does not target a specific objective.
26	Eliminate. This question is too difficult. Introductory astronomy students have not needed to think about how gravity changes with distribution of mass.
27	Completely re-work this question to get at the same idea, but be more explicit by using familiar objects, like the Sun, Moon, Mars, and a person on Earth’s surface.
28	Do not refer to “surface gravity.” Instead use weight as a way to compare the strength of the gravitational force on Earth and Saturn. Re-word choices to be similar in length.
29	Eliminate. There are too many nuances to interpreting this question. It tests an understanding of projectile motion more than it tests student understanding of gravity.
30	Eliminate. The choices are too similar in meaning. Also, interviews suggest students do not read carefully, interpreting 0.25% as 25%. So, some interviewees indicated that they did not believe the questions stem, since they had never noticed feeling lighter when changing elevations, such as hiking up a mountain or riding in a plane. There are more effective ways of testing how the strength of the gravitational force diminishes with distance.
31	Eliminate. The choices are targeting too many different misconceptions at once. Also, a student may want to choose more than one of the available answers, so guessing is more likely.

APPENDIX L

CONCEPT INVENTORY VERSION 2

Gravity Concept Inventory
Version 2

Please answer the following questions about Newtonian gravity to the best of your ability. This questionnaire is for research purposes only, so your performance will NOT affect your grade in any way, and your participation is optional. Your student ID number will be used for matching your responses to other data, but your information will be destroyed after the study is over.

SELECT THE SINGLE BEST CHOICE

1. Besides the force of gravity, which of the following factors hold(s) us to the Earth's surface?
   a. Air pressure from the Earth's atmosphere.
   b. Forces from the Earth's spinning motion.
   c. Magnetism from the Earth's magnetic field.
   d. More than one of the above factors.
   e. No other factors; gravity is the only one.

2. 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?
   a. The force would become two times weaker.
   b. The force would not change.
   c. The force would become two times stronger.
   d. None of the above is correct.

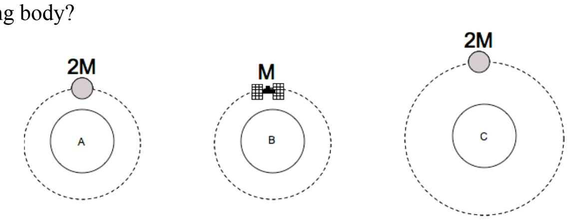
3. Two astronauts are floating in space very far away from any planets or stars. What is the direction of the gravitational force that they experience, if any?

   a. Toward each other, because there is a gravitational force between them.
   b. Away from each other because they are pulled by distant planets and stars.
   c. They experience a gravitational force, but its direction cannot be determined.
   d. They do not experience a gravitational force because there is no large object nearby.

4. Astronauts appear weightless in their Earth-orbiting spacecraft because:
   a. They have escaped Earth's gravity.
   b. There is no air in the spacecraft.
   c. They are moving at the same speed as their spacecraft.
   d. The spacecraft's rocket engines counteract gravity.

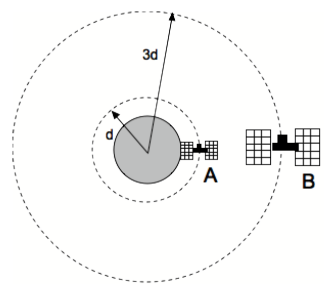
5. The factors that most affect the gravitational force between two objects are[i]:
   a. size and distance
   b. mass and size
   c. density and distance
   d. mass and distance
   e. density and size

6. Planets A, B and C are identical. A and C each have a moon of mass 2M orbiting them, while B has an artificial satellite of mass M orbiting it, as shown in the diagram. Which planet has the strongest gravitational interaction with its orbiting body?



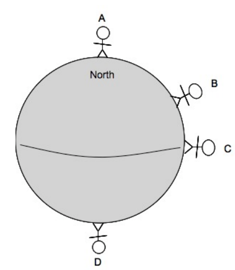
   a. Planet A
   b. Planet B
   c. Planet C
   d. All the same.
   e. Not enough information given to determine.

7. Satellites B is *three* times more massive than A, but orbiting the planet at *three* times the distance. Compare the force of gravity between each satellite and the planet. The diagram is not to scale.



   a. A experiences a stronger force than B.
   b. B experiences a stronger force than A.
   c. A and B experience the same force.
   d. A and B experience no force at all.

8. Assuming the Earth is perfectly round and each person is the same mass, which person standing on Earth's surface experiences a stronger force of gravity?



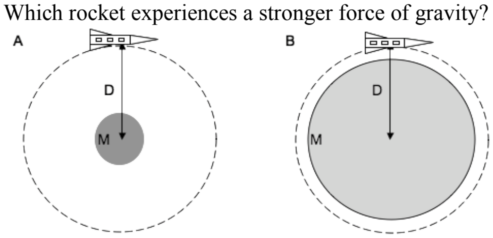
   a. Both A and D
   b. B
   c. C
   d. D
   e. All the same

9. Why does the Earth exert a gravitational force on objects on its surface?
   a. It has an atmosphere
   b. It is very dense.
   c. It has a magnetic field.
   d. It has mass.
   e. It rotates.

10. Suppose the Sun shrank in size but its mass remained the same so that it was compacted much more densely. How would the gravitational force experienced by the Earth change[ii]?
   a. Earth would experience a weaker gravitational force.
   b. Earth would experience the same gravitational force.
   c. Earth would experience a stronger gravitational force.
   d. It cannot be determined with the given information.

11. What is the relationship between a planet's gravity and its atmosphere?
    a. Atmosphere represents a boundary where objects can experience a gravitational force.
    b. Atmosphere causes objects on the surface to experience a stronger gravitational force.
    c. The gravitational force holds the particles in the atmosphere close to the planet.
    d. None of the above is correct.

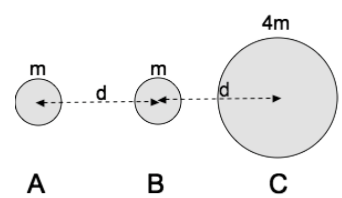
12. Each planet below has the same mass M and an identical rocket orbiting at a distance D. The darker the planet, the denser the material. Which rocket experiences a stronger force of gravity?
    a. A
    b. B
    c. Both the same
    d. Not enough information given to determine.



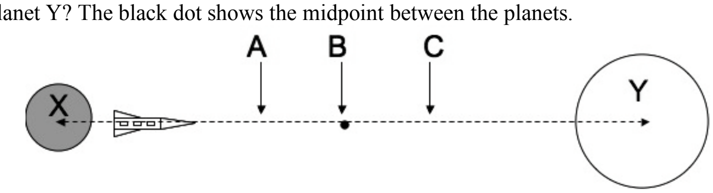
13. The gravitational force between two objects is approximately zero:
    a. when another object lies in-between the two original objects.
    b. when the distance between them becomes extremely large.
    c. at a location halfway between the two objects.
    d. when they are balanced in a stable orbit around one another.
    e. when the two objects are not moving.

14. The following diagram shows two smaller planets of mass *m* and one larger planet of mass *4m*, each separated by distance *d*.

    The gravitational force *by Planet C on Planet A* ______ that *by Planet B on Planet A*.
    a. is less than
    b. is equal to
    c. is greater than
    d. cannot be compared with the given information to.

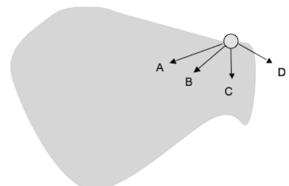


15. Assume that planets X and Y have the same mass, but planet X is made of a denser material. At which point would a rocket experience a gravitational force from planet X that is **equal to** that from planet Y? The black dot shows the midpoint between the planets.



    a. At point A, because planet Y is bigger.
    b. At point B, because planets X and Y have the same mass.
    c. At point C, because planet X is more dense.
    d. The rocket does not experience a gravitational force with X or Y, because it is in space.
    e. None of the above is correct.

16. Consider an irregular-shaped asteroid of constant density shown below. Which arrow best represents the direction of the gravitational force on the ball?



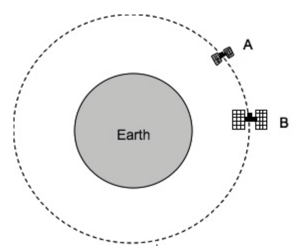
   a. A, because it points to the center of mass of the asteroid.
   b. B, because it points to the geometric center of the asteroid.
   c. C, because it points straight down to directly below the ball.
   d. D, because it points to the edge nearest the ball.

17. An astronaut standing on the Moon's surface has a pen in his hand and releases it. What happens to the pen[iii]?
   a. It falls to the surface at a slower rate than it would on Earth.
   b. It falls to the surface at the same rate as it would on Earth.
   b. It falls to the surface at a faster rate than it would on Earth.
   c. It doesn't fall and floats where it is.
   d. It doesn't fall and slowly drifts away.

18. Does an Earth-orbiting satellite or the Moon experience a greater force of gravity from the Earth?
   a. The satellite, because it is much closer to Earth.
   b. The moon, because it is much more massive.
   c. They experience the same force, because they both orbit.
   d. Neither experiences a force, because they are in space.
   e. Not enough information given to determine.

19. Theoretically, how far away from the Earth do you have to go to feel *zero* gravitational force from the Earth?
   a. Outside the atmosphere
   b. The orbit of the Moon
   c. The orbit of Mars
   d. Outside the solar system
   e. An infinite distance

20. Satellite B is *three* times more massive than Satellite A. Which satellite experiences a stronger gravitational attraction with the Earth?

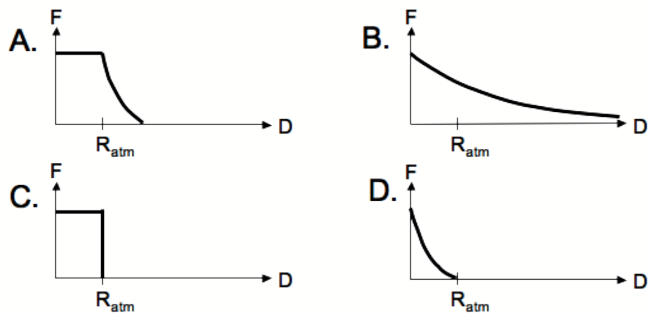


   a. Neither, since they are in space.
   b. Neither, they are at the same distance.
   c. A, since it is easier to move.
   d. B, since it has more mass.
   e. None of the above is correct.

21. Which of the following would make you weigh half as much as you do right now[iv]?
   a. Take away half of the Earth's atmosphere.
   b. Take away half of the Earth's mass.
   c. Take away half of the Earth's magnetic field.
   d. Make the Earth spin half as fast.
   e. Move Earth twice as far from the Sun.

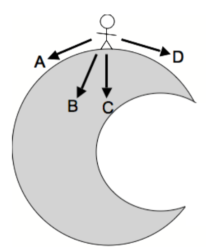
22. The reason why your weight at the cloudy surface of Saturn is similar to that on Earth is that:
    a. Saturn has the same mass as Earth, and your weight depends only on the planet's mass.
    b. Saturn's large size would cause you to weigh more, but the rings counter the effect.
    c. Saturn is much more massive, but on its surface you are further from its center.
    d. Saturn is much more massive, but it is also much farther from the Sun.
    e. None of the above is correct.

23. Which graph best represents the gravitational force F that a rocket experiences as it travels a distance D away from the surface of the Earth? Note that R_{atm} represents the edge of the atmosphere.



24. Which one of the following statements about gravity and magnetism is true?
    a. Gravity cannot exist without magnetism.
    b. Gravity is not affected by magnetism.
    c. Magnetism increases a planet's gravity.
    d. Magnetism cancels out a planet's gravity.

25. A person is living on a planet where a huge hole has been carved out. Which arrow best represents the direction of the gravitational force he experiences?

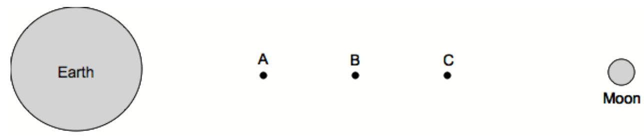


    a. A, because it points along the surface with the most mass.
    b. B, because it points directly to the center of mass.
    c. C, because it points straight down to the center.
    d. D, because it points to the hole and he would fall in.

26. Of the objects listed at right, which experiences a gravitational force from the Earth?
    a. None of them
    b. Person only
    c. Moon and Person
    d. Moon, Person, and Sun
    e. All of them

Sun
Moon
Person
Mars

27. In the diagram below, at which point would a spacecraft traveling from the Earth to the Moon experience a gravitational force from the Earth that is equal to the force it would feel from the Moon? Note that B is at the midpoint between the Earth and the Moon[v].



a. A, the spacecraft is closer to Earth and Earth is more massive.
b. B, the spacecraft is exactly halfway between Earth and the Moon.
c. C, the spacecraft is further from Earth and Earth is more massive.
d. All points are the same, since the spacecraft is in space.

Demographic questions

28. How many astronomy and/or physics courses have you taken before this one?
a. 0
b. 1
c. 2
d. more than 2

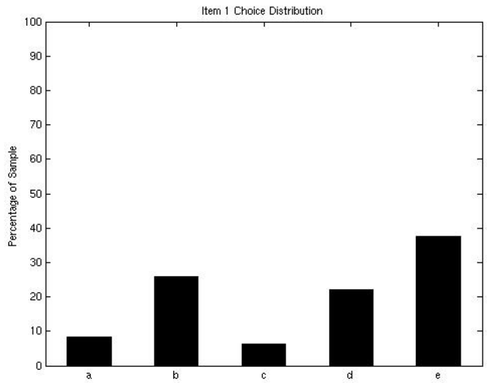
29. What is your gender?
a. Male
b. Female

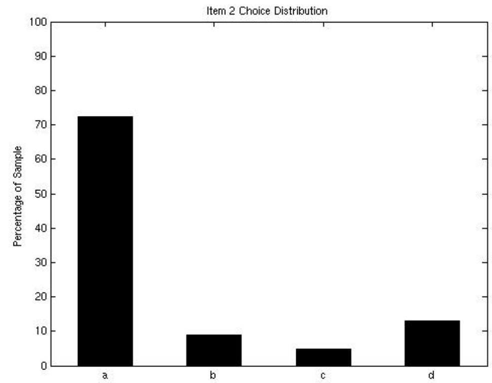
---

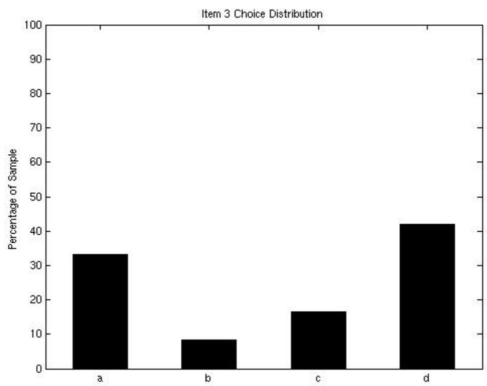
[i] Adapted from E.E. Prather's assessment.
[ii] Adapted from a question by Philip Blanco.
[iii] Adapted from J. Dostal's Master Thesis.
[iv] Adapted from the TOAST.
[v] Adapted from Hudgins, D.W. & Prather, E.E. 2007, *Astronomy Education Review*, 5(1).

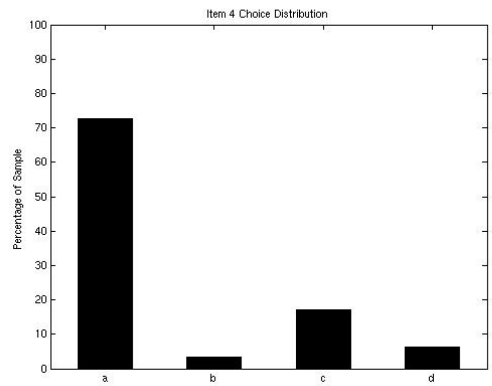
APPENDIX M

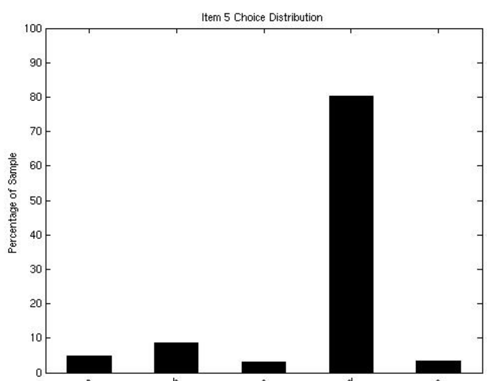
VERSION 2 STUDENT RESPONSES

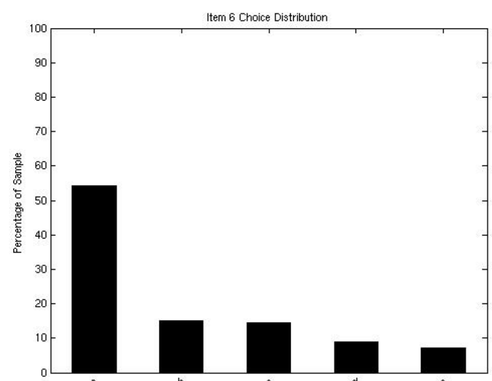

Item 1 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

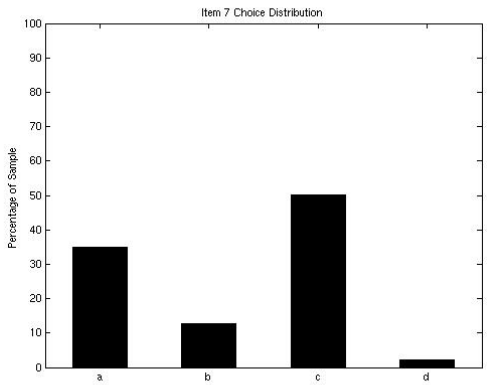

Item 2 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

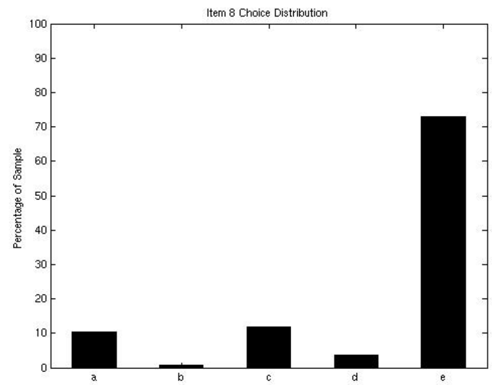

Item 3 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

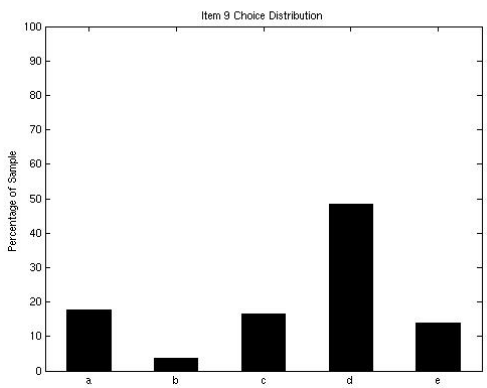

Item 4 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

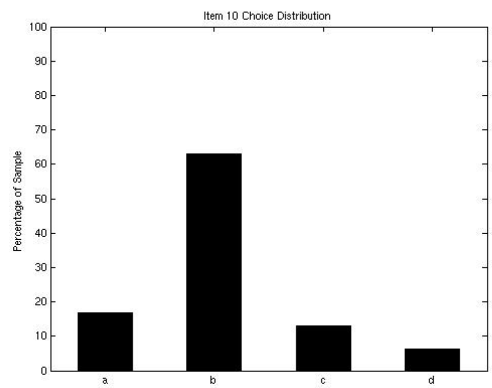

Item 5 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

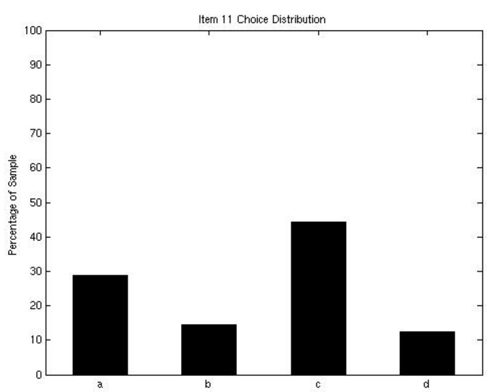

Item 6 Choice Distribution
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

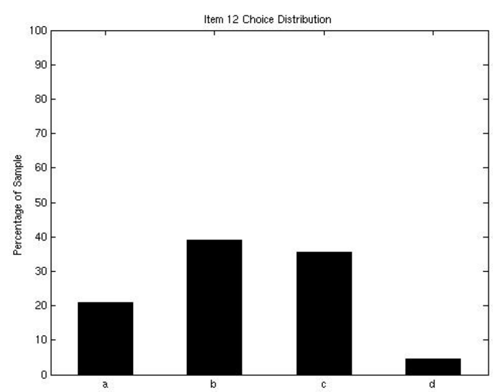

Item 7 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

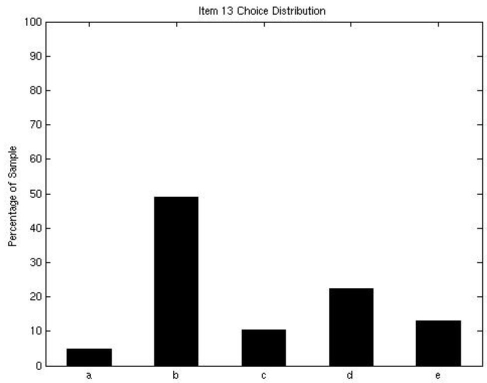

Item 8 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e

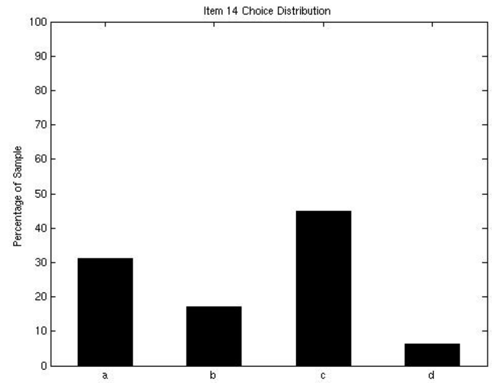

Item 9 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e

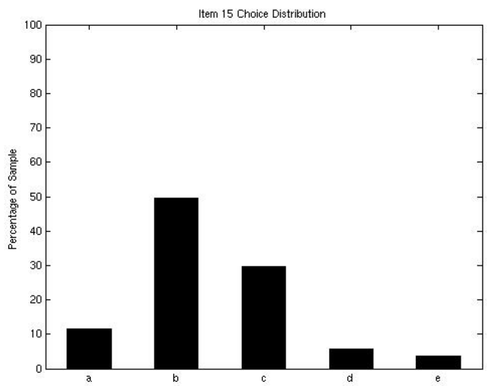

Item 10 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

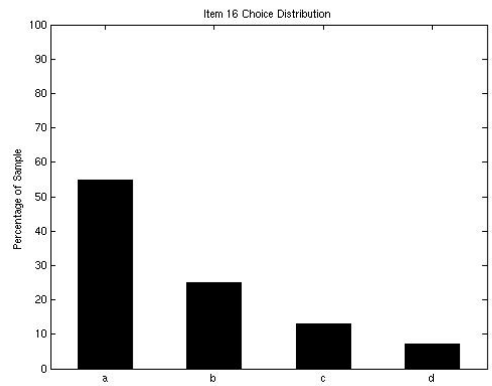

Item 11 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

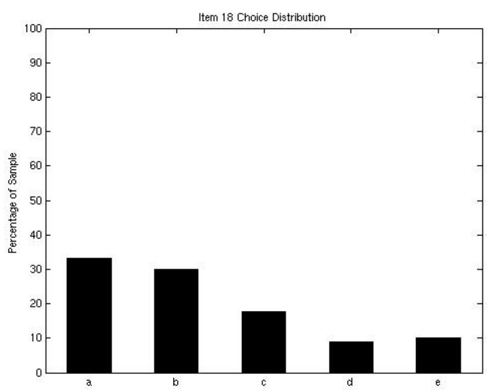

Item 12 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

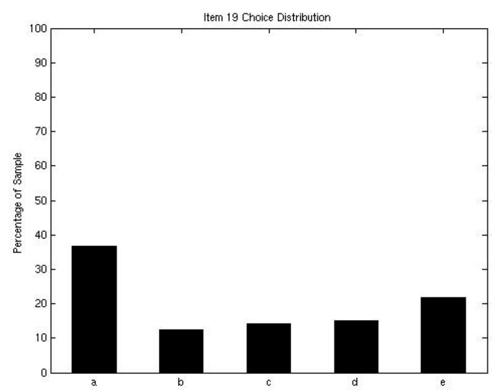

Item 13 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

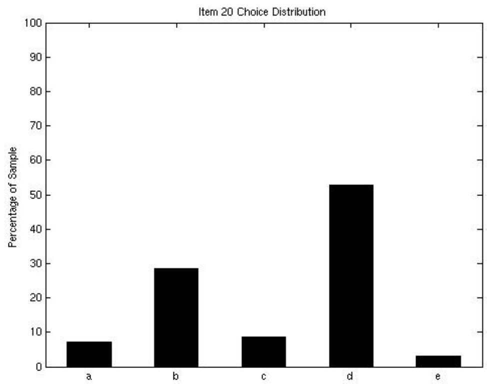

Item 14 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d

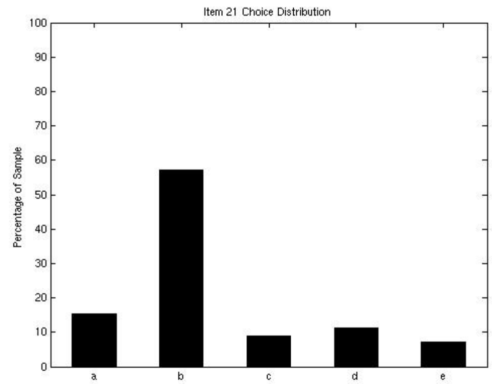

Item 15 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

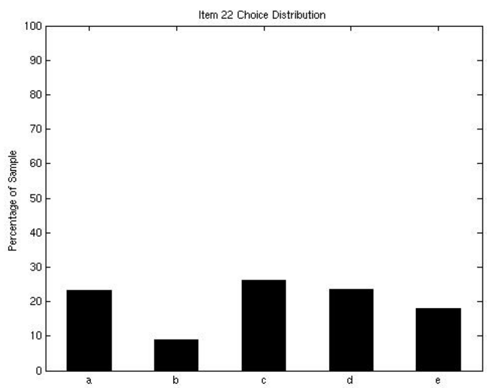

Item 16 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d

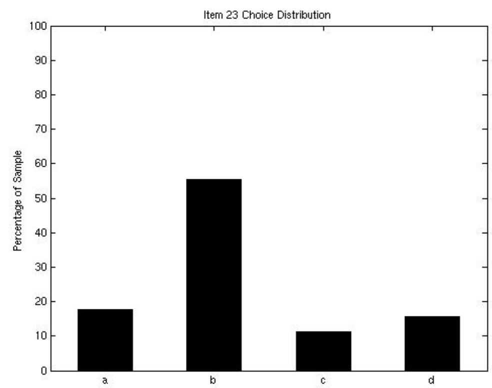

Item 18 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

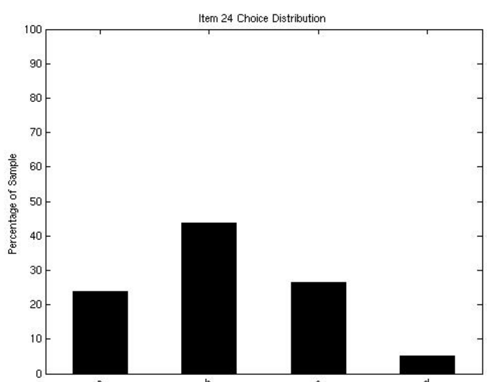

Item 19 Choice Distribution
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

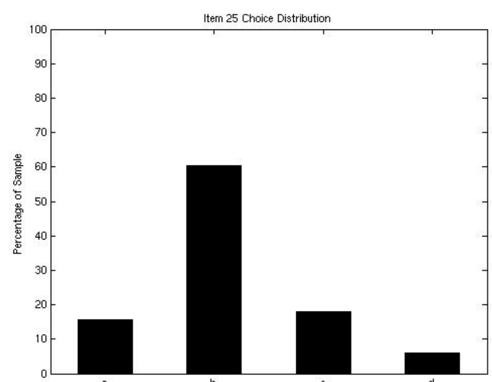

Item 20 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e

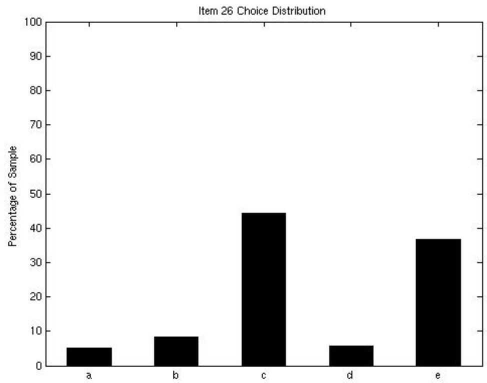

Item 21 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e

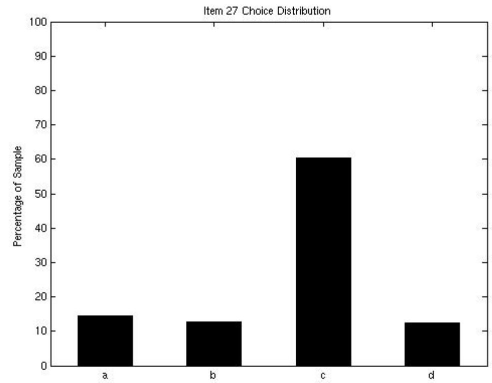

Item 22 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e


Item 23 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d


Item 24 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d


Item 25 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d


Item 26 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d
e


Item 27 Choice Distribution
Percentage of Sample
100
90
80
70
60
50
40
30
20
10
0
a
b
c
d

APPENDIX N

STUDENT INTERVIEW NOTES FOR VERSION 2

Student Information	Interview Notes
F11-01 Took physics before	- ‘Learned in class’ that the strength of gravity is directly proportional to distance. - Question 7 – more mass is exactly compensated by more distance – effects cancel. - Every mass has a gravitational force, no matter how small - Doesn’t understand how moving at same rate as spacecraft leads to weightlessness. Chooses this answer by process of elimination. This is not how he would answer the question, but he doesn’t see his answer as a choice. - Don’t need to consider the mass of the person on Earth. Uses test taking strategies – if it’s not in the question it’s not a variable. - At one point says that gravity is only affected by mass, but at others considered both distance and mass. - Believes that the atmosphere is of constant density and that the ozone layer is the casing that hold the atmosphere in, and that spacecraft must ‘break free’ from this boundary - At first gets confused by Newton’s laws, but when I phrase is as a by/on, he remembers and changes his reasoning - Confusion about orbiting – gravity pulls satellite around.
F11-02 Physics major, took Physics in high school, in Physics now, learned about gravity in Physics yesterday	- Immediately wrote the gravitational force equation meticulously, including all subscripts. She used this first and foremost in approaching problems, even when proportional reasoning wasn’t necessary. It took more time to apply the formula than to apply the concept. - She seemed to apply ‘radius’ both correctly and incorrectly. For the astronauts in space, she drew the r-vector from a point between them to one of them. Maybe she thinks that it is always measured from the center of mass of the system? - She made a few algebra errors that caused her to get the wrong answer, but she had the right idea. False negatives. - It was interesting to see her reason through Q18. She wrote the equation and labeled variables as ‘very large’ to make approximations and comparisons. She realized there was still not enough information. - She didn’t think gravity was affected by magnetism b/c she hadn’t heard of anything like that before.
F11-03 Music Education, No previous physics or astronomy courses	- Mixing of Forces – rotation affects gravity, but not sure. Also maybe magnetism. Notes that all choices in Q9 are true so tries to use process of elimination. - Misapplication of Distance: doesn’t know the proportionality. It seems reasonable that it would be a direct

	relationship. - Q4 False Positive: Thought that astronauts orbiting are not experiencing gravity, but wasn't sure if that's why they were weightless. - Only says that there is no gravity in space for questions that directly confront that. - Used observation and experience to deduce that gravity is the same at different locations on Earth. - Constantly double checking her logic on each question to determine if mass or density is more important. She remembers learning about black holes recently and how dense they are. - Chooses certain answers just b/c they are a middle ground among the choices, such as choice 'c' in Q17. - Direction: Ball would roll off of an irregular asteroid in Q16 - Chooses to pay attention to distance or mass, not the combination. - Can measure gravity by its effects: person in Q25 would be pulled along side with most mass b/c he is constrained to the surface.
F11-04	- Mixing of Forces: Earth's spinning motion helps keep us on Earth, and it causes people at the equator to feel a stronger force of gravity, but she did seem to realize that it was separate (Earth would still have gravity without spinning). She said that the difference in force between a person at the equator and a person at the pole would be hardly noticeable. - Thinks that atmosphere is important for the effects of gravity, but recognizes that there is a gravitational force past the atmosphere (chose the correct graph and referenced in many times). She concludes that there must be a difference between force and pull to reconcile her confusion about why objects orbiting past the atmosphere don't fall down. - Was very sure that magnetism was separate from gravity. She compared gravity and magnetism on other planets to back up her choice. - Misapplication of mass: She understands that both masses matter, and that there is no minimum mass, but she thinks there has to be a *difference* in mass to cause a force. Ex. One object has to be 5x more massive than the other object. - The direction of the force is from the smaller object to the larger object. But if you had two Earths side by side, the force would go both ways. There would be a push and a pull, because each planet would be able to "hold its own ground" - Believed there was a one-to-one relationship with mass and

	distance. She used the ratio idea often, she just didn't realize that the distance was squared. - She considered the distance to the 'middle' of each planet in all questions except Q12. She was just going with her gut reaction to how the picture looked.
F11-05	- Mixing of Forces: Air pressure and rotation keeps us on Earth. Compares to spinning pulling you away from something. References how high people can jump at higher altitudes from exercise science? People at the poles would experience a stronger force of gravity b/c they are not spinning. - Boundary Model: Atmosphere is the boundary - Misapplication of distance: Chose the direct (one-to-one) relationship between mass and distance – just assumed. Also measured distance from surface at first, then changed. - Relies on a gut feeling to answer Q12. Appears to have some confusion about mass vs. density – says that they are pretty much the same thing. Used process of elimination of Q13, saying that the other choices just don't make sense. - Misapplication of Mass: Astronauts in space do not experience gravity b/c they are outside the atmosphere, but the Moon, Mars, and the Sun do b/c they are so large – minimum mass?
F11-06	- Immediately wrote gravitational force equation, except had a 'g' for gravity rather than an 'F'. I couldn't tell if he mislabeled, or if he knew the difference between force and acceleration - Clearly understand proportional reasoning. - Said 'greater force so greater acceleration' for a pen falling on Earth compared to Moon. - I'm surprised that he did not get question 18. He jumped to "same force b/c both are orbiting" and didn't pay attention to choice 'e'. Q20 causes him to rethink Q18. He's reminded that distance matters. - Q20 drew a free body diagram. Knew the force for gravity, and drew other arrows but couldn't figure out what caused those forces. Was uncomfortable with an 'unbalanced' free body diagram. He had forgotten about centripetal force. - Q22 considers 'a', but is pretty sure Saturn is more massive.
F11-07 Took high school physics	- Initially deduces that rotation is necessary for gravity based on hearing about a rotating space station, but then later figures out that rotation isn't necessary, it just creates 'artificial gravity'

	- Knows the 1/r^2 rule, and applied it properly a few times on his own, and then after prompted in a few questions where he forgot - Couldn't remember the formula exactly, but mass matters - Gravity equal for each person on Earth just seems reasonable - Consistent in reasoning that density does not matter - Reasoned that a smaller satellite experiences more gravity. Acknowledged that more mass definitely means more force, but claimed that there is a difference between the force and the gravity that the satellite 'feels' – still stuck to the orbital indicator model - Very solid understanding of the role of the center of mass as the center of gravity
F11-08 University Studies	- Mixing of Forces: magnetic field keep us on Earth's surface, but also chose that it's just mass for another question. - Mixture of correct scientific model and Misapplication of distance: knows that force diminishes with distance. Chose correct graph, but does not know the proportionality. Chooses to stick with one-to-one relationship between mass and distance to stay consistent, since it makes the most sense. - Very weak gravity outside of Earth's atmosphere. With discussion, decides speed has something to do with weightlessness. - Chose mass and distance as the most influential factors. Assumes that both masses matter.
F11-09 Finance	- Misapplication of Distance: knows that distance diminishes, and chooses a direct relationship just because that's what is given. - Misapplication of mass: Two astronauts are too small to create gravity. An object has to be about the size of the Moon at least. - Boundary Model: The atmosphere is the boundary for gravity for astronauts in an orbiting space ship. It's gradual. - Factors: size and distance, changes to mass and distance. - Distance to Sun matters: Initially said that people at the equator experience a stronger gravity because they are closer to the Sun, then realized that they are sometimes farther from Sun, so changed answer to people at the poles. - Mixing of Forces: Earth has gravity because it has an atmosphere. If it didn't have atmosphere, we wouldn't have gravity. A pen on the Moon wouldn't fall and would float away because the Moon doesn't have an atmosphere. The Moon would eventually move away and leave the pen

	behind. Has no idea how magnetism relates to gravity. Only a person experiences a gravitational force from Earth. Take away half of Earth's atmosphere to take away half of Earth's gravity, and half of Earth's mass. Chose graph A of Q23 – only when you lose atmosphere do you lose gravity. - Thinks hard about the interplay of size and density for Q12. Decides he has no idea and chooses D. - Orbital Indicator Model: The force of gravity is zero when objects are balanced in a stable orbit. There is no force if they are not falling down or not crashing in to each other. Chose D on Q14 because you would need to know the orbital dynamics of the planets to deduce their gravity. - Density vs. Mass: Chose 'c' on Q15 because planet X is denser. Mixed up the difference between center of mass and geometric center. - Q27 not sure if Moon has gravity. Thought the rocket would just crash into the Moon b/c it is in the way, not because the Moon is pulling it. Ended up choosing D, all points the same.
F11-10	- Misapplication of mass: Understand that just gravity holds us to Earth, but then says that also the mass of Earth as those were separate ideas. The object with the larger mass matters. - Misapplication of distance: knows that greater distance means smaller gravity, but doesn't know mathematical form. The size of planets matter (more than just mass and distance) – larger planets have stronger gravity. - Mixing of forces: astronauts escape gravity when they escape magnetic field. It's a cloud of force. People at the poles would feel stronger gravity b/c of magnetism. Magnetism increases a planet's gravity - Wondered if your weight on Saturn is same as on Earth b/c Saturn is so far from Sun. - Measure gravity by its effects: Only a person and the Moon are experiencing a gravitational force with Earth.
F11-11 Music Education	- very unsure, but was able to pinpoint where her confusion comes from - didn't think air pressure affected gravity, but wondered in rotation and magnetism would. - A direct relationship for distance just seems reasonable - Understood the diminishing force with infinitely large distances, so considers that external stars and planets DO pull on astronauts, but still chose that they experience a force with each other. BUT, she also said that it's true that astronauts are weightless because they escaped Earth's

	gravity - More mass means more gravity - Chose that all people on Earth will experience the same gravity b/c of experience, no particular theory, it just seems reasonable - As the questions started asking more about density, she decided to change her framework for answering questions away from mass as a primary factor and more toward density. This was a conscious decision. She also based this on recently learning about black holes and how dense they are and how much gravity they have. - Felt like the ball would fall off the asteroid and fall 'down'. It's just a guess, but it feels right. - Doesn't know what would happen to a pen on the Moon, chose floating b/c it seemed like a middle ground - A person on the planet with a hole carved out would be drawn to the side with the most mass. Maybe the center of mass is pulling him, but he's constrained to the surface, so that's the direction the force must point. It must point in the direction of possible motion.
F11-12 Business Accounting Major Took Physics in High School	- Mostly correct mental model, very consistent - Mixing of Forces: Remembers learning about centrifugal forces in high school and wonders if that affects gravity. Pretty sure that magnetism doesn't affect gravity just because he never remembers having to consider magnetism when talking about gravity. - Misapplication of Distance: Used an inverse law rather than an inverse square law for distance. - Correct scientific model: Force drops off to infinity. Both masses matter. It's a ratio between mass and distance.
F11-13 Business Management 01919824	- Mixing of Forces – magnetism and rotation keep us on Earth - Misapplication of mass - Astronauts in space are not massive enough to gravitationally interact. Would have to be the size of a mountain maybe. - Boundary Model – atmosphere holds gravity in. - Misapplication of distance – from surface, consistently. - Correct scientific model – understands that both masses matter. References knowledge of the existence of gravitational force equation, but can't recall what it is. Pretty sure it doesn't have magnetism in it. - Reconciles the mass and distance model with the Mixing of Forces model by saying that magnetism creates the gravitational field, but mass and distance determine its strength.

	- Draws on previous knowledge – imagines some of the planets to be the Sun, Earth, and Moon - Was confused by Q14, with planet B interfering with gravity - Learns from the test – uses Q15 to rethink Q12 distance from center vs. surface. - Moon doesn't have gravity, so pen wouldn't fall. When we go to the Moon, astronauts have to get their own force of gravity from the air pressure in their spacesuits. - She realized that there was inconsistency in her logic when confronted with reconciling her idea of no gravity in space with the Moon orbiting earth, but didn't feel the need to resolve it. - Was at least consistent in saying that Earth's gravity is only within the atmosphere – only a person would feel it, not Moon, etc. Although, it's interesting that she doesn't consider this at all when confronted with questions that don't address it. Her understanding is clearly a result of her interaction with a question.
F11-14 Undecided Took high school physics	- Very confused, said that most of her answers were just guesses at the 60%/40% level, but had reasons for choosing answers, even if she wasn't confident. - Mixing of Forces: Was hooked on the term 'escape velocity' and took it to mean escaping Earth's atmosphere. But was confused because she recognized that orbiting implies that it is still experiencing a gravitational force. Decided to say that rocket engines. Magnetism increases a planet's gravity because magnetism is tied to the atmosphere and you need atmosphere for gravity. - Once she decided that mass and distance were the primary factors in Q5 she really stuck with it throughout the test. I wonder if she would have gotten some questions wrong had she not seen Q5. - Said "Sure" to my suggestions. - Said that a pen would slowly drift away from the Moon b/c she saw it in movies. She agreed that there was no gravity on the Moon when I suggested that's what her answer implies simply because it goes with her answer, but she said she didn't think that was right. - Chose the correct answer for Q26, all objects are in our galaxy. I don't know if this is a false positive or not.
F11-15	- Mixing of Forces: Earth's spinning motion keeps us on the surface. Said that magnetism increases a planet's gravity because magnetism and gravity both come from the same source: rotation. The center of mass is what everything

	rotates around and therefore what everything is attracted to. - Didn't seem to interpret questions in a consistent way: He got the graph of the gravitational force correct, but he also believed that atmosphere affects gravity. Or rather, atmosphere represents the boundary where most of Earth's gravitational force is concentrated. - Orbital Indicator: An object must be falling toward or moving around another object if it experiencing a gravitational force. Objects that are just sitting there are not experiencing a force. Although, he also said that the gravitational force is zero when objects are balanced in a stable orbit, so this is contradictory. Two astronauts in space would feel each other's gravity, so they would start to orbit each other and then spiral in. Wonders if objects can cancel each other's gravity. Tries to use a "maximum orbit" argument to determine how far an object can leave Earth and still feel a gravitational pull. - Distance: He knew that the force goes to zero at infinity, but he also thought there was a way to have zero gravitational force closer. - Scientific Model: Mass and distance are the two factors that determine gravity. Mass is the thing that determines the force, and distance determines the range of the force. Size and density matter indirectly. - Didn't know if he should add the two masses or subtract them, but knew that both mattered.
F11-16	- Mixing of Forces: Rotation creates gravity. Earth's rotation keeps us on the surface. You could reduce your weight in half by reducing the spin of Earth by half. Your weight on Saturn is similar to Earth because Saturn spins slower. Referenced space stations as needing to spin to have gravity. The Moon has weaker gravity because it rotates slower. - Scientific Model: Mass and distance are the two factors that affect the gravitational strength. Gravity diminishes with distance, but doesn't know the relationship. - Misapplication of Distance: Decides that the proportions are one-to-one. - Misapplication of Mass: Two astronauts do not feel a force because there are no stars or planets nearby. - Confused about Direction: Objects are pulled straight down. Thought that the equator was the "center" of Earth, and noted that objects do not fall "toward the equator." - Boundary Model: The atmosphere is like a boundary for gravity, because once you get past the atmosphere you're not

	going to experience gravity in the same way. Satellites would fall if they were just placed in the sky, but they are able to orbit outside the atmosphere. Although, they must be experiencing some gravity, because they eventually fall back down. The Moon must be at a location where there is just enough gravity to keep it from moving closer. You must have to go beyond the orbit of the Moon to experience zero gravity.
F11-17 Fish & Wildlife Management Took Physics in High School	- Scientific Model: Mass and distance are the only two factors that affect the gravitational force. She very clearly did NOT have the Mixing of Forces misconception. She explicitly stated that Magnetism doesn't affect the gravitational force, although both forces can affect a single body, so it might *look* like they were related. - Misapplication of Distance: Wrote the gravitational force equation, but forgot to square distance, so she got most of the proportionality stuff wrong. - Misapplication of mass: Did not consistently use the gravitational force equation in terms of proportional reasoning. She thought that it was the "difference" in mass that mattered, but it seemed like she meant the ratio of masses rather than subtraction. Either way, she did not realize that it was really a multiplication until very late in the test, and she didn't go back to correct her previous mistakes. On Q22, said your weight depends only on the planet's mass. - She was confused about how a bowling ball and feather could fall at the same rate and be feeling different forces. She used this to try to understand Q20, and eventually settled into the correct answer. - Deduced that there must be no relationship between a planet's gravity and its atmosphere because Venus has a very different atmosphere, but about the same mass.

APPENDIX O

MODIFICATIONS FROM VERSION 2 TO VERSION 3

Item	Modifications
1	No changes.
2	Too difficult, because the sample population of students did not focus on the inverse square law in class. Move this question to another section of the test. When it is next to Item 1, there are two "none of the above" correct answers in a row. Boldface "between the Earth and the Sun." Interviews suggest students don't read carefully. Interviews suggest distractor choices are not functioning as intended. Change the choices to be only in relation to "weaker" (2x weaker, 4x weaker, 8x weaker, and 16x weaker).
3	No changes.
4	Too difficult. Interviews suggest that the choices are worded awkwardly. There were some false negatives. Maybe the word "falling" would cue the students that do understand orbiting to answer correctly more than "moving". Change the choices to be more direct: either astronauts are experiencing a gravitational force from Earth or not, OR the spacecraft has a force that counters the pull from Earth.
5	Too easy. Interviews corroborate this, and once students answer this question they use it throughout the rest of the survey. This question is an easy way for students to "learn from the test," and so it distracts from their true understanding of gravity. It should be eliminated.
6	Re-word without mathematical symbols. Give the planets shading. Choice 'e' isn't very good distractor. Replace with "Both A and B".
7	Put the satellites on opposite sides of the planet. Remove "diagram not to scale" – it's not necessary. Eliminate choice 'd' – it is not a plausible distractor.
8	Add "from Earth" to the end of the question. Change it to one person walking along the Earth, rather than different people. Specifying that "each person is the same mass" is sort of a hint to students. Choices 'b' and 'd' are not good distractors. Eliminate them.
9	Choice 'b' is not a good distractor. Eliminate it and add the choice "more than one of the above."
10	Choice 'd' is not a good distractor. Eliminate it.
11	This question needs to be eliminated. While it is of adequate difficulty and discriminatory power, it does not target a single objective. If a student chooses 'a' then he may be working from the Boundary Model. If he chooses 'b' then he is having issues with Mixing of Forces; if he chooses 'c' he understands that there is no minimum mass requirement for the gravitational force, and if he chooses 'd' what he thinks is unclear.
12	Have a stronger contrast in the darkness of each planet. Boldface some of the key information. Some students do not read carefully. Choice 'd' is not an effective distractor. Eliminate it. Determine if it is fair to ask this question given that students are not expected to

understand Gauss's Law.

13 Good difficulty and discriminatory power. Choice 'a' is not an effective distractor. Eliminate it. Add the choice "more than one of the above."

14 This question is very difficult for the MSU population, but will likely be appropriate when wider pilot testing is implemented for Version 3. Choice 'd' is not an effective distractor. Eliminate it.

15 Good difficulty and discriminatory power. Choices 'd' and 'e' are not effective distractors. Eliminate them.
Clarify that the distance is measured from the center of each planet.

16 Good difficulty and discriminatory power.
The phrases "geometric center" and "center of mass" might be jargon. Make the difference between these terms clearer.

17 This question was not included in the statistical analysis because it had two choice 'b's. Regardless, hardly any students chose 'b', so both 'b's are probably bad distractors. Eliminate the choice "it falls at the same rate as it would on earth."

18 Eliminate it. Definitely too difficult. Even students that understand everything about the gravitational force equation struggle with it. At least one professor was not sure how to approach this question. He ended up calculating it and determined that the Moon experiences more force from Earth than a satellite.

19 Eliminate. There is a better way to probe understanding of how the force from Earth diminishes as a rocket leaves.

20 This question is at a perfect difficulty level for this population, but not a good discriminator. Eliminate choice 'e' – very few chose it.

21 No change. This is a good question.

22 Students definitely struggle with this question. Interviews suggest that some students simply do not believe that Saturn's mass is the same as Earth's. Consider eliminating it. But, it does target a single objective: student ability to combine effects of mass and distance. Replace it with an imaginary scenario that asserts an equal force of gravity on the surfaces of a small planet and a large planet and requires students to determine *how* this could be the case.

23 Eliminate variable shorthand and use words – some students are confused by "R_{atm}."Add another graph that is a combination of choices A and B. Each distractor is effective.

24 Choice 'd' is not an effective distractor. Eliminate it.

25 Good difficulty and discriminatory power. However, this item might not be fulfilling a particular objective. There are better ways to probe for student understanding of direction. Eliminate it.

26 Choice 'a' is not a good distractor. Eliminate it.
Choice 'd' was not a very good distractor either, but keep it for now. It makes the choices more complete. The table is unnecessary. Include the choices in the question stem.

27 Good difficulty and discriminator power. All distractors are effective.
Clarify that B is the midpoint between the centers of the Moon and Earth.
Reduce the text by eliminating the stems of the choices.

APPENDIX P

CONCEPT INVENTORY VERSION 3

Newtonian Gravity Concept Inventory
Version 3

Please answer the following questions about Newtonian gravity to the best of your ability on the bubble sheet provided. Use a #2 pencil, and make sure your name and answers are in the appropriate positions. *Select the single BEST choice.*

DO NOT START THE SURVEY UNTIL YOU ARE TOLD TO DO SO

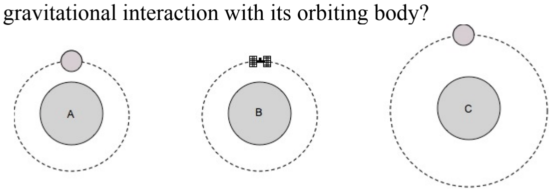
PLEASE DO NOT WRITE ON THIS SURVEY

1. Besides the force of gravity, which of the following factors hold(s) us to Earth's surface?
   a. Air pressure from Earth's atmosphere.
   b. Forces from Earth's spinning motion.
   c. Magnetism from Earth's magnetic field.
   d. More than one of the above factors.
   e. No other significant factors; only gravity.

2. Two astronauts are floating in space very far away from any planets or stars. What is the direction of the gravitational force that they experience, if any?

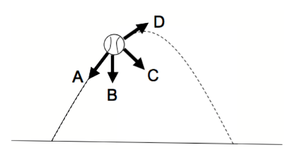
   a. Toward each other, because there is a gravitational force between them.
   b. Away from each other because they are pulled by distant planets and stars.
   c. They experience a gravitational force, but its direction cannot be determined.
   d. They do not experience a gravitational force because there is no large object nearby.

3. 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. Planet A
   b. Planet B
   c. Planet C
   d. Both Planets A and B
   e. All the same



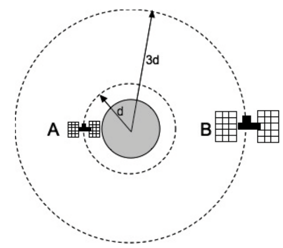
4. If the Moon were located *four times farther* from Earth as it is now, the gravitational force between Earth and the Moon would become…
   a. Two times weaker.
   b. Four times weaker.
   c. Eight times weaker.
   d. Sixteen times weaker.

5. 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?
   a. A
   b. B
   c. C
   d. D



6. An astronaut floating in her Earth-orbiting spacecraft …
   a. experiences no gravitational force from Earth.
   b. still experiences a gravitational force from Earth.
   c. experiences a force from the spaceship that counters the gravitational force from Earth.

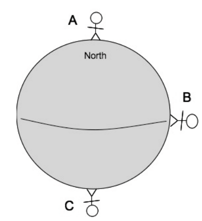
7. Satellite B is *three* times more massive than A, but orbiting the planet at *three* times the distance. Compare the force of gravity between each satellite and the planet.



a. A experiences a stronger force than B.
b. B experiences a stronger force than A.
c. A and B experience the same force.

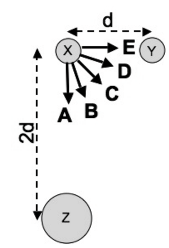
8. Why does Earth exert a gravitational force on objects on its surface?
a. It has an atmosphere
b. It has a magnetic field.
c. It has mass.
d. It rotates.
e. More than one of these.

9. Assume Earth is perfectly round and that a person walks from the North Pole to the South Pole. At which position does the person experience a stronger force of gravity from Earth?



a. A only
b. B only
c. Both positions A and C
d. Each point is the same

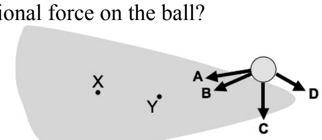
10. Three planets are arranged as shown in the diagram. Planets X and Y each have mass *m* and Planet Z has mass *2m*. Planet X is a distance *d* away from Planet Y and a distance *2d* away from Planet Z. Which arrow (A-E) best represents the direction of the *total* (net) gravitational force on Planet X?



a. A
b. B
c. C
d. D
e. E

11. If the Sun shrank in size but its mass remained the same so that it was compacted much more densely, the gravitational force between Earth and the Sun would…
a. become weaker.
b. remain the same.
c. become stronger.

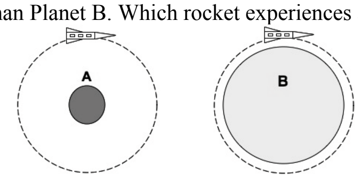
12. For the asteroid shown below, X represents the center of mass and Y represents the geometric center. Which arrow best represents the direction of the gravitational force on the ball?



a. A, because it points to X.
b. B, because it points to Y.
c. C, because it points beneath the ball.
d. D, because it points down the slope.

13. The gravitational force between two objects is approximately zero…
    a. when the distance between them becomes extremely large.
    b. at a location halfway between the two objects.
    c. when one is balanced in a stable orbit around the other.
    d. when the two objects are not moving.
    e. More than one of the above is correct.

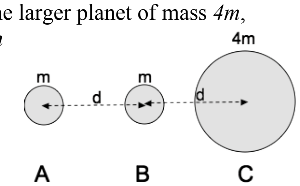
14. Each planet below has the *same mass* and an identical rocket orbiting at the *same distance* from its center, but Planet A is *denser* and *smaller* than Planet B. Which rocket experiences a stronger force of gravity?



    a. A
    b. B
    c. Both the same

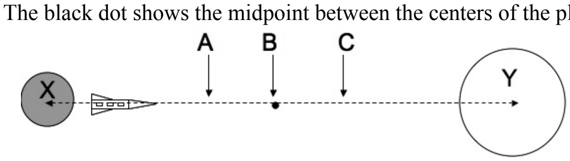
15. Which one of the following statements about gravity and magnetism is true?
    a. Gravity cannot exist without magnetism.
    b. Gravity is not affected by magnetism.
    c. Magnetism increases a planet's gravity.

16. The following diagram shows two smaller planets of mass *m* and one larger planet of mass *4m*, each separated by distance *d*. The gravitational force *by Planet C on Planet A* is ______ that *by Planet B on Planet A*.



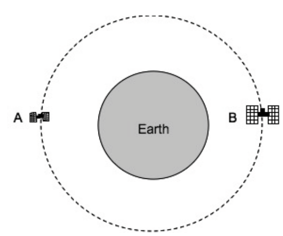
    a. is less than
    b. is equal to
    c. is greater than

17. Assume that planets X and Y have the same mass, but planet X is made of a denser material. At which point would a rocket experience a gravitational force from planet X that is **equal to** that from planet Y? The black dot shows the midpoint between the centers of the planets.



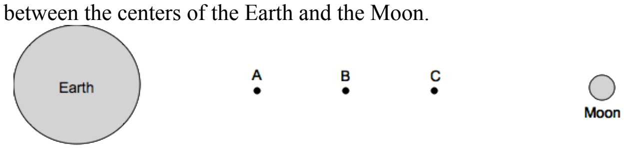
    a. At point A, because planet Y is bigger.
    b. At point B, because planets X and Y have the same mass.
    c. At point C, because planet X is more dense.

18. Satellite B is *three* times more massive than Satellite A. Which satellite experiences a stronger gravitational attraction with Earth?



    a. Neither, since they are in space.
    b. Neither, they are at the same distance.
    c. A, since it is easier to move.
    d. B, since it has more mass.

19. In the diagram below, at which point would a spacecraft traveling from Earth to the Moon experience a gravitational force from Earth that is equal to the force it would feel from the Moon? Note that B is at the midpoint between the centers of the Earth and the Moon.
    a. A
    b. B
    c. C
    d. All points are the same

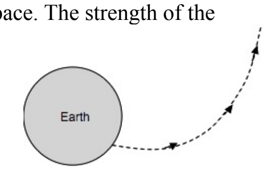


20. An astronaut standing on the Moon's surface has a pen in his hand and releases it. What happens to the pen?
    a. It falls to the surface at a slower rate than it would on Earth.
    b. It falls to the surface at a faster rate than it would on Earth.
    c. It doesn't fall and floats where it is.
    d. It doesn't fall and slowly drifts away.

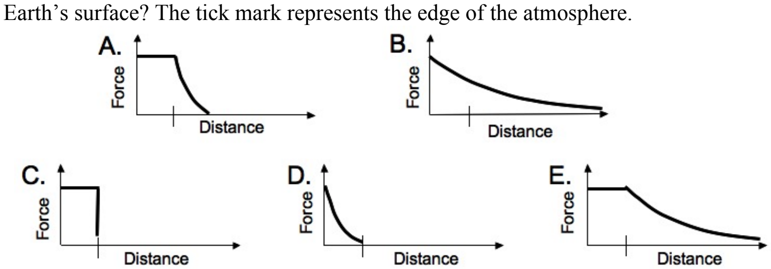
21. Which of these objects – **Sun, Moon, Person, Mars** – experience a gravitational force from Earth?
    a. Person only
    b. Moon and Person
    c. Moon, Person, and Sun
    d. All of them

22. Which of the following would make you weigh half as much as you do right now?
    a. Take away half of Earth's atmosphere.
    b. Take away half of Earth's mass.
    c. Take away half of Earth's magnetic field.
    d. Make Earth spin half as fast.
    e. Move Earth twice as far from the Sun.

23. A rocket is launched from Earth and it travels farther and farther into space. The strength of the gravitational force it experiences from Earth will…
    a. drop to zero immediately after it leaves the atmosphere.
    b. eventually level out to a constant value greater than zero.
    c. eventually be exactly zero at a location in our solar system.
    d. get smaller and smaller, but will never reach zero.

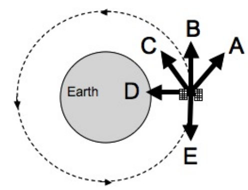


24. Which graph best represents the gravitational force that a rocket experiences as it travels away from Earth's surface? The tick mark represents the edge of the atmosphere.



25. For your weight to be the same on both Planets A and B …
    a. Planet A must have a denser atmosphere because it is smaller.
    b. Planet A must have the same mass as Planet B.
    c. Planet B must be more massive and rotate faster.
    d. Planet B must be more massive because your location is farther from its center.

26. An Earth-orbiting satellite is shown at right. Which arrow, if any, best represents the direction of the gravitational force that it experiences?
    a. A
    b. B
    c. C
    d. D
    e. E



Consent and Demographic questions
Your responses in this survey can help with a research project on student understanding of gravity. Your participation in the research project is voluntary and will not affect your grade in any way, and your personal information will be kept confidential.

27. Do you consent to allowing your answers to this survey to be analyzed for the research project?
    a. Yes
    b. No

28. What is your age?
    a. 17 years or younger
    b. 18-30 years
    c. older than 30 years

29. What is your college major (or current area of interest if undecided)?
    a. Business
    b. Education
    c. Humanities, Social Sciences, or the Arts
    d. Science, Engineering, or Architecture
    e. Other

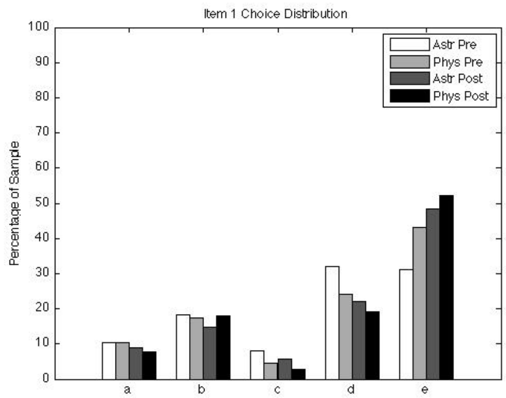
30. How many astronomy and/or physics courses have you taken before this one?
    a. 0
    b. 1
    c. 2 or more

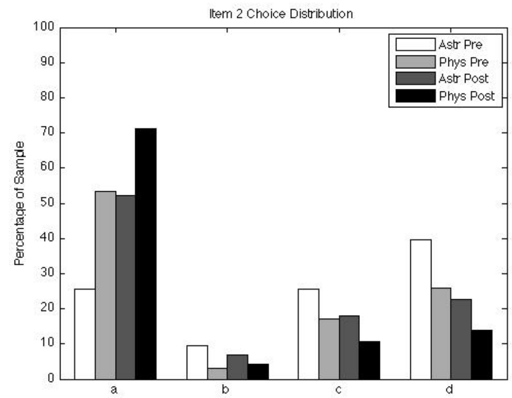
31. What is your gender?
    a. Male
    b. Female

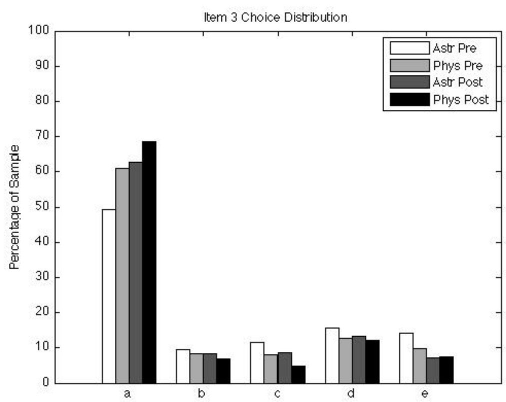
32. Which best describes the community where you live?
    a. Rural
    b. Small town
    c. Suburban
    d. Urban
    e. Not in the USA

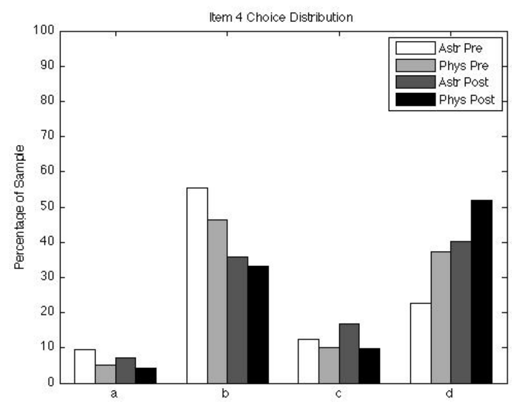
APPENDIX Q

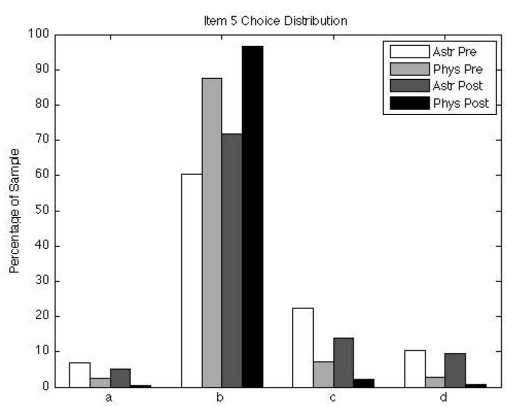
VERSION 3 STUDENT RESPONSES

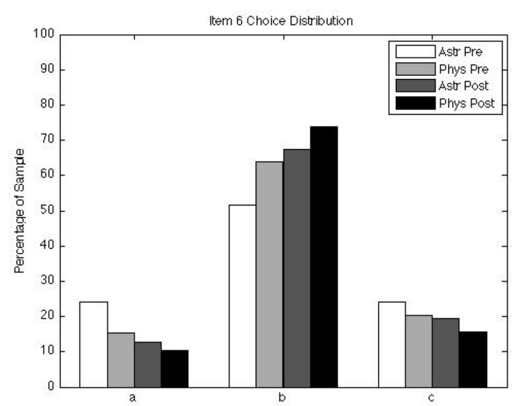

Item 1 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c
d
e

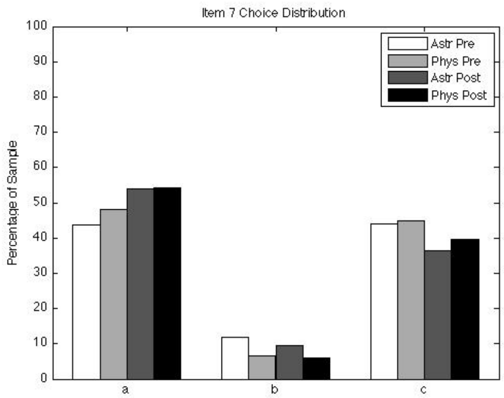

Item 2 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c
d

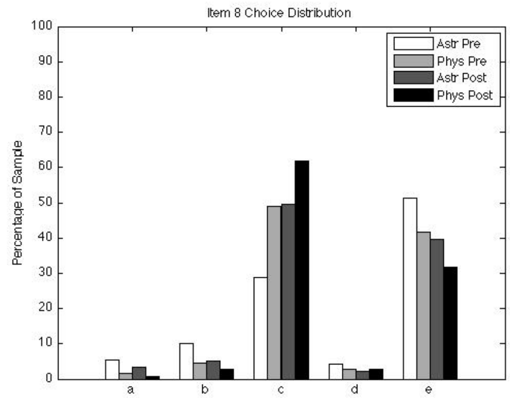

Item 3 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c
d
e

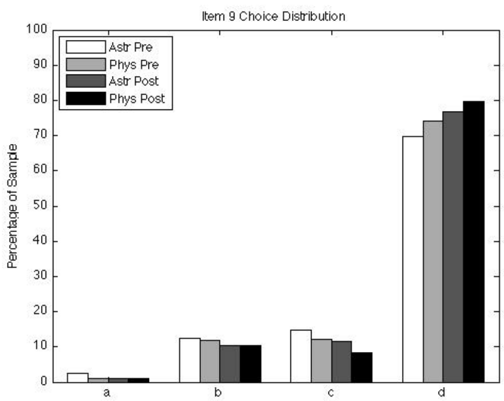

Item 4 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c
d

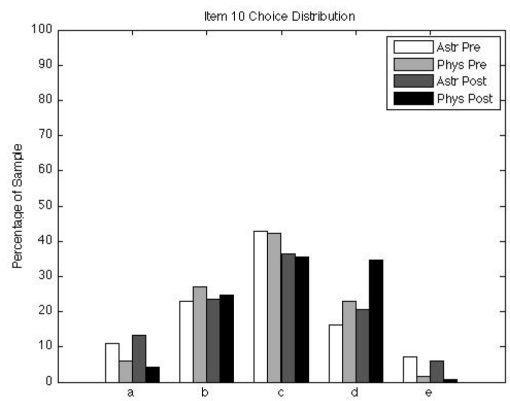

Item 5 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c
d

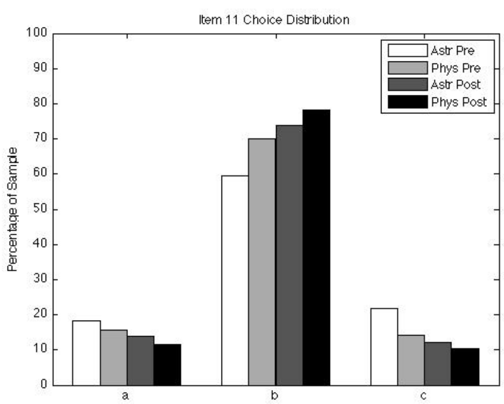

Item 6 Choice Distribution
Percentage of Sample
Astr Pre
Phys Pre
Astr Post
Phys Post
a
b
c

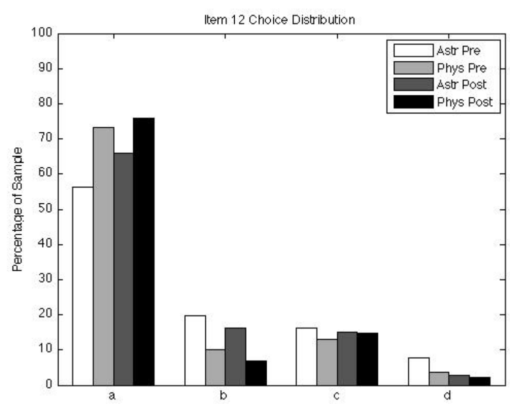

Item 7 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c

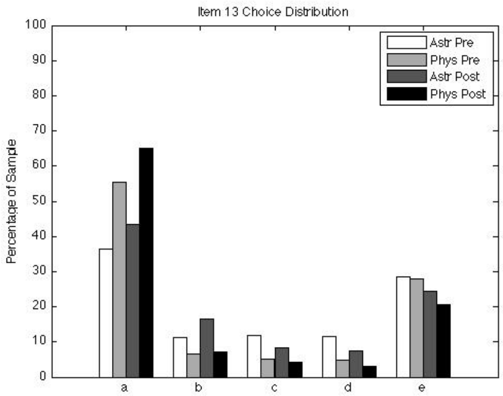

Item 8 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d
e

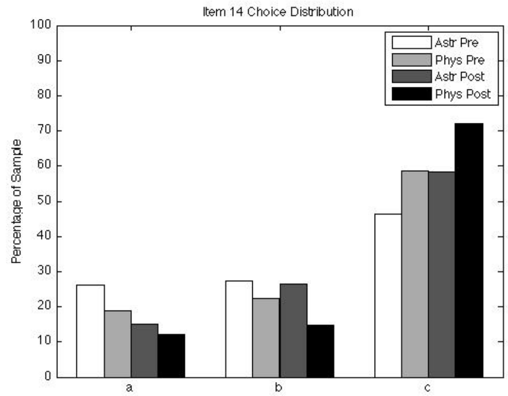

Item 9 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d

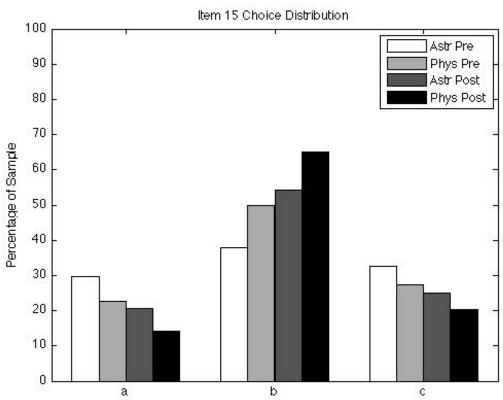

Item 10 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d
e

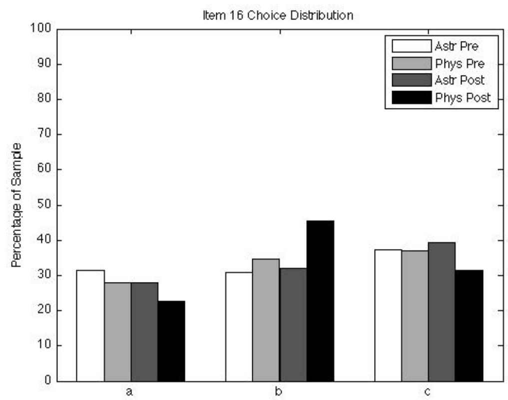

Item 11 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c

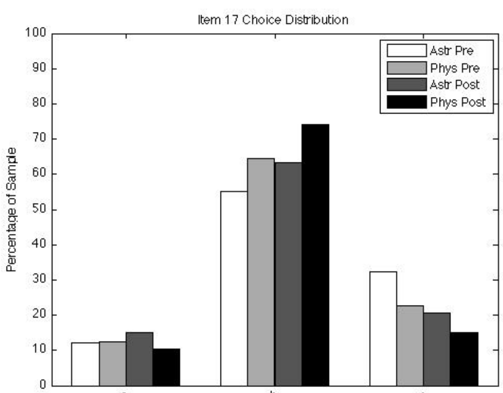

Item 12 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d

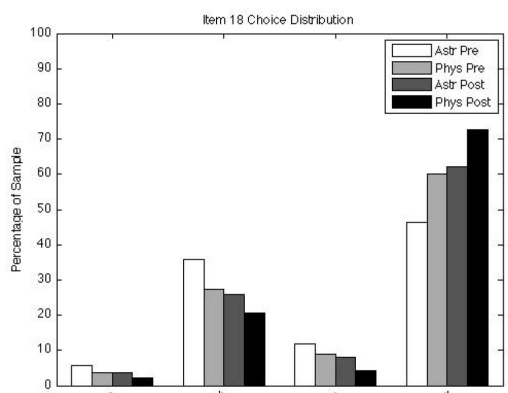

Item 13 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d e

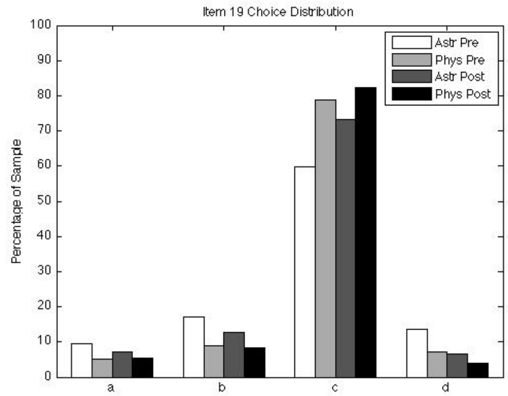

Item 14 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c

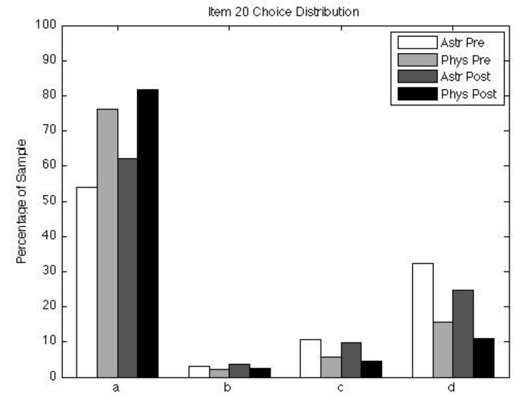

Item 15 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c

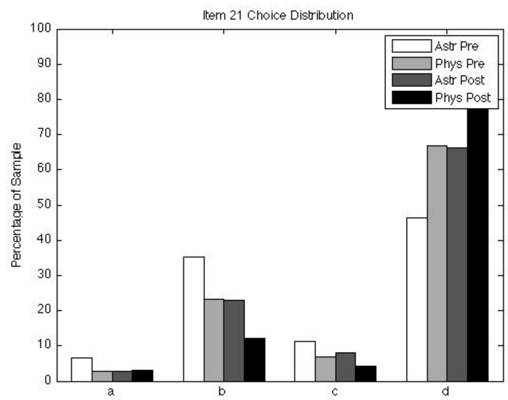

Item 16 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c

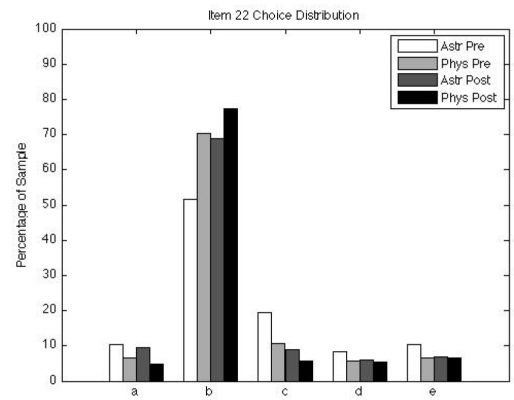

Item 17 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c

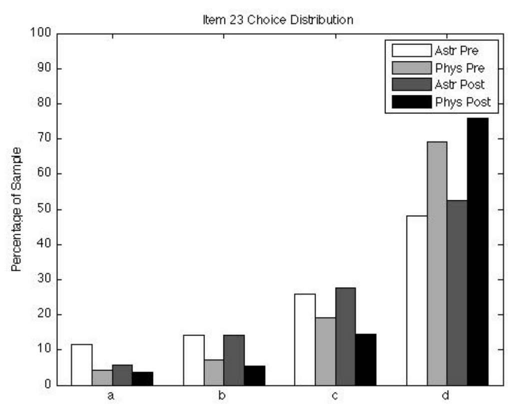

Item 18 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0 10 20 30 40 50 60 70 80 90 100
a b c d

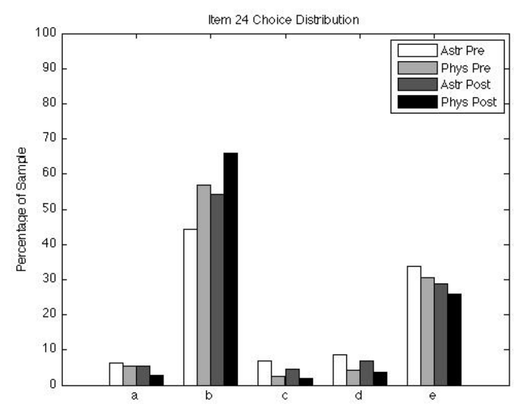

Item 19 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d

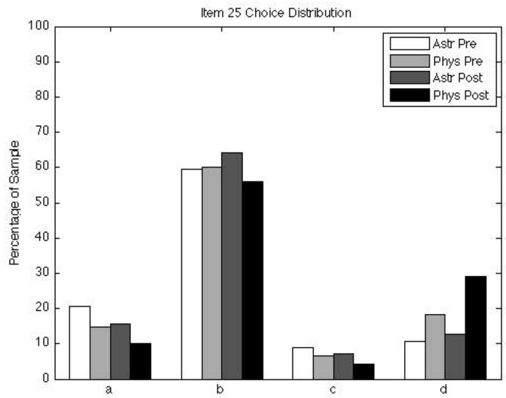

Item 20 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d

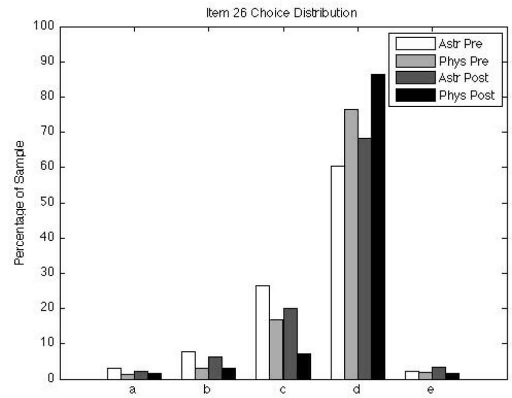

Item 21 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d


Item 22 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d
e


Item 23 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d


Item 24 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
a
b
c
d
e


Item 25 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d


Item 26 Choice Distribution
Astr Pre
Phys Pre
Astr Post
Phys Post
Percentage of Sample
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e

APPENDIX R

ASTRO 101 STUDENT INTERVIEW NOTES FOR NGCI VERSION 3

Student Information	Interview Notes
S12-01	- Application of the Scientific Model – consistently used the inverse square law and reasoned correctly with both mass and distance variables. - Boundary Model – Thought that there was both a minimum mass required for something to exert gravity (two astronauts in space are not heavy enough to "warp spacetime" – maybe a little but not noticeable) and that there is some place where the force drops to zero (although this location would be far away, not sure if it's in the solar system or beyond). - Mostly no mixing of forces: Realized that although an object can experience a magnetic and gravitational force at the same time, gravity itself is not affected by magnetism. He realized that the centripetal force at the equator might make a person feel lighter, but he did not separate the concepts of "apparent weight" from "gravitational force" and chose that persons A and C would experience a stronger force. Positive that atmosphere has no effect on gravity. - Correct direction: to the center of mass, "down"
S12-02 Psychology No previous courses	- Mixing of Forces: Was pretty sure that no other factors affected the force of gravity. If anything, rotation might. If magnetism did, he didn't know how. Thought about it and decided atmosphere doesn't affect it. Consistently chose only gravity, but with some hesitation each time. - Recognized the similarity between certain questions. Tried to stay consistent. Tried to learn from the test – searching for a question that would confirm his way of thinking. - Misapplication of distance: Q25 helped him decide that distance is "distance from the surface". This was consistent with him saying that the force from the Sun would decrease if the Sun shrank, and Q14 (rocket ships orbiting). - Misapplication of the force law: Recognized when a question required mathematical manipulation, but he couldn't remember the equation. He understood the basic direct/indirect nature of mass and distance, but had no idea that distance should be squared – ex. Though 16x weaker for Q4 was 'absurd.' He settled on a 1:1 ratio and used that consistently, acknowledging that he wasn't sure that was right, but it seemed better than nothing. Also tried to rely on the visuals, ex: Q10.

S12-03	- Was very consistent. No mixing of forces model. Knew that gravity diminished steadily with distance, but got caught up on Q24 because she thought about when rockets leave Earth, suddenly astronauts are floating. - Misapplication of force law: Understood the basic direct/indirect nature of mass and distance, but didn't know the distance was squared. It seemed like a compromise to have these effects cancel each other out with a 1:1 ratio. - Knew that things fall slower on the Moon because she's seen videos of astronauts, but didn't really understand why. Figured out that it's probably b/c of less mass, but didn't seem confident. - Said that Q14 is visually tempting to choose B, but she knows the force is the same. - Knew that distance was measured from the center of mass, not surface.
S12-04 Forensic Anthropology No previous courses	- Mixing of Forces: Q1: Thought air pressure held us to Earth's surface and magnetism, even though she said she didn't know humans were magnetic, she stuck with her answer. Q8: magnetic field and atmosphere. Not rotation, since that would fling you off. - Realized that Earth's gravity always pulls things straight down (Q5 and Q26), but on Q2 said that the direction would depend on where they came from. She said something about objects staying in motion in space (maybe newton's first law?) and decided there was a relationship between speed and direction of gravity - Misapplication of Force Equation: Knew that mass and distance mattered. When confronted with Q4, automatically knew 16x weaker because she's seen questions that ask for her to square something, but she didn't realize what quantity she was squaring (distance). Later she decided that the mass probably matters more than distance, so maybe that was the quantity that gets squared. She knew there wasn't a 1:1 relationship - Q11: False positive. Decided that if the Sun shrank the whole solar system would shrink accordingly, and with all the same proportions gravity would stay the same. - Misapplication of distance: Was very unsure about how force changes with distance. Knew that it doesn't drop to zero right after the atmosphere (satellites are able to orbit there), but it didn't seem right that it would never get to zero, so she chose that it would get to zero in our solar system on Q23, and realized that Q24 was very similar,

	deciding that the force would drop off *after* the atmosphere (your weight on Mt. Everest is the same as at the bottom of the ocean, or maybe we just don't notice?). Q21: Maybe sometimes we exert a force on Mars when we're close to it in our orbit, but maybe not all the time.
S12-05	- Thought air pressure held us to Earth's surface. At first thought rotation and then realized he was thinking about a spaceship, and that would fling you out. - Knew proportionality without knowing the equation. He knew distance was a more influential factor than mass because of the square, but he said the force between the Moon and Earth would just be reduced by four if the Moon moved 4x further. - He probably understood more than what his answers would indicate because he did not read the questions carefully and rushed through the test. I had to ask him to explain his reasoning on almost every question. He said that the gravitational force would level out, but ignored the "to a constant value" part. Chose the correct graph in Q24, and when I asked why he did not stay consistent, he realized he read Q23 wrong. - Magnetism is completely separate from gravity. - Understood direction completely.
S12-06 AguCon 1 physics class in high school	- Mixing of Forces: Spinning motion and air pressure affect gravity. If you took away some of the atmosphere, gravity would be reduced (although not sure of the proportionality) - Scientific Model: Knew that mass (not density) and distance (from the center) affect gravity. Knew that there is no minimum mass or maximum distance. - Proportionality: Did not know about the inverse square law. Thought it reasonable that mass and distance would have a 1:1 relationship, and used this consistently, but not confident. - Overall felt very unsure about the test. Felt most sure about the pen on the Moon problem. - Distance: Eventually the distance levels out to a constant value rather than constantly decreasing. At first chose E for Q24 and then changed answer to B because gravity drops as soon as you leave Earth (i.e. its surface).
S12-07 Cell Biology/ Neuroscience 3 previous courses	- No mixing of forces: Only gravity holds you to Earth, magnetism is a separate force. Atmosphere doesn't matter. - Scientific Model: Knew the equation really well, and consistently tried to manipulate it with the brute force method by plugging in example numbers, but struggled

	with proportional and conceptual reasoning. Sometimes she got the right answer and sometimes she made a mistake calculating. - Misapplication of distance: Sometimes measured distance to surface and sometimes knew it was center of mass. - Misapplication of mass: Astronauts would not experience a force from each other. Knew a pen would fall to the surface of the Moon, but mentioned there might be something that is too small of a mass so that it would just float away. With further questioning and reflection, she concluded it must be zero mass and she didn't know of anything that was zero mass. - Density doesn't matter. - Mars wouldn't feel a force from Earth. Described how we know Moon, person and Sun experience gravity and couldn't think of a way to know. I think she was looking for a visible effect, but I'm not sure.

APPENDIX S

PHYS 101 STUDENT INTERVIEW NOTES FOR NGCI VERSION 3

Student Information	Interview Notes
F12-01 Construction Engineering & Technology 3 previous classes	- Scientific Model: Had a very solid understanding of gravity. Got every question correct once he remembered that distance has an inverse square relationship. - The only question he struggled with at all was item 25. He read all the choices and drew relative distances on the diagram, which made him feel more confident.
F12-02 Pre-Med Cell Biology	- Mixing of forces: Air pressure and magnetism keep us to Earth's surface - Misapplication of Distance: only inverse relationship - Direction: chose arrows that point in the direction that the object would move toward for items 5 and 26 - Misapplication of Mass: Density is the most important factor and mass is secondary.
F12-03 Geology Took Physics I and Astro 101 in Spring 2012	- Mixing of Forces: Item 1: Air pressure helps hold us to Earth's surface. But then in item 24 says that atmosphere has an effect because it adds to Earth's mass. Reasons that there is no definite edge to the atmosphere so gravity cannot suddenly drop there. - Misapplication of Distance: only inverse relationship. - Refers to F=ma for gravity. Seems to think that the acceleration is gravity. Applies this to Item 13: if something is not moving it's not accelerating and therefore it has no gravity. But, for Item 20 reasons that a smaller force implies a smaller acceleration, determining that the pen must fall slower. - Struggled with Item 10. Re-read the question a few times. Chose D because the closer distance of Y outweighs the larger mass of Z.
F12-04 Cell Bio and Neuroscience, Psychology minor	- Misapplication of Distance: Used the correct inverse square law but called it "exponential." Q24: Different masses affect the rate that gravity falls off with distance. A rocket ship is small so the force it experiences would fall off more slowly. - Scientific Model: No minimum mass (two astronauts would attract each other). Correctly used proportionality with distance and mass (Q7). Correctly uses mass vs. density - Q5: Chose B because it was "perpendicular to the plane of the Earth" - Orbital Indicator: Q3 wondered if Planet C would have stronger gravity because it could hold the moon so far

	away. Q21: Mars and Earth are not orbiting so there is not a gravitational force between them. - Mixing of Forces: Rotation: Q9 – a person at the equator would need to experience a stronger force of gravity since it's spinning. Magnetism: Q15 always regarded magnetism as separate but it's possible that it does contribute, so chose C, magnetism increases gravity - Q22: Changing distance to Sun would change Earth's gravity
F12-05 Secondary Education General/Broad Field Science 1 previous course	- Mixing of Forces: Q1: Spinning motion keeps us on Earth (guess). Q8: the choices are true, so more than one (poor test taking strategies). Q15: gravity cannot exist w/out magnetism - Threshold Model: Q2: astronauts in space wouldn't feel attracted to each other. - Scientific Model: Q3: Closer and more massive means more force. - Misapplication of Distance: Force and distance are directly proportional (Q4). Equal changes in distance cancel changes in mass (Q10). Q13: chose combination E for halfway and large distance. - Q19: was confused because thought that Earth's pull would always be stronger than the Moon's since Earth is bigger. - Q22: distance from Sun matters. References the SPOT show slide about the gravity on Mars.
F12-06 Microbiology 1 previous course	- Misapplication of Distance: Direction relationship (Q4). - Scientific Model: Mass and distance are the primary factors that determine the strength of gravity. Gravity is not affected by magnets. - Inconsistent: Q1 only gravity holds us to Earth, but then for Q8 said that the atmosphere is part of the reason Earth exerts a gravitational force, because the atmosphere is created by gravity – is this circular logic? - Equal changes in distance cancel changes in mass (Q16).
F12-07 Construction Engineering and Technology 2 previous courses	- Scientific Model: No minimum mass (Q2 astronauts are pulled toward each other). Astronauts are flying at the same rate as their spacecraft so they float. - Inconsistent: Uses a one-to-one relationship between mass and distance (Q4), with equal changes in distance cancel changes in mass (Q10). Bust, asserts distance plays a bigger role than mass in Q7 and Q16.
F12-08 Exercise Science	- Mixing of Forces: Rotation: Spinning *causes* gravity (Q1), the poles of Earth have stronger gravity because of rotation

2 previous courses	(Q9), Q22: spinning faster would hold you down more. Magnetism: decided gravity cannot exist without magnetism because something must hold objects in place, but it was somewhat of a guess - Threshold Model: Minimum mass for gravity (Q2 – astronauts wouldn't feel force because far away from everything). - Inconsistent misapplication of distance and mass: Remembered that distance involved a square, but took the square root and chose 2x weaker for Q4. Then said equal changes in distance cancel changes in mass (Q7, Q16). Then, said equal changes in mass outweigh changes in distance (Q10). For Q19, chose A, had heard that the pull from the Moon is greater than Earth, but then for Q20 said that a pen would just float on the Moon. - Direction: always straight down (even for Q12). - Scientific Model: Was able to separate the effects of mass and density (Q11, Q17, Q14). Q21: all objects experience a gravitational force with Earth. Steady decline of force with distance (Q23, Q24). - Mentioned superposition.
F12-09 Construction Engineering and Technology	- Mixing of Forces: not sure if spinning holds us to Earth. Other forces definitely act on us, but wasn't able to separate those forces from gravity. Q8: Not sure which choices he would choose, but felt that it was more than one of those factors. - Threshold Model: Minimum mass for gravity (Q2 – astronauts would feel force from distant objects, but direction can't be determined). - Confused gravity for acceleration. Q3: gravity is a constant so A and B would be equal. Said weight equals mass times gravity. Got Q9 correct, but mentioned that, "gravity is a constant on the surface of Earth." - Misapplication of distance: One-to-one relationship (Q4), but acknowledge that it's probably an r^2 relationship, but decided to stick with 4x weaker. Equal changes in mass cancel changes in distance (Q7 to stay true to consistency, Q10, Q16). Gravitational force would be zero halfway between two objects - Scientific Model: Was able to separate the effects of mass and density (Q11, Q14, Q17). Steady decline of force with distance (Q23, Q24). Q15: Decided that magnetism is probably not affected by gravity. Pen would fall slower on Moon. Q19: Rocket would need to be closer to Moon to

	experience equal force with Earth. Got all direction questions correct. - Extra notes: Q25: Gravitational force is the result of weight of a planet rather than its size. At the end of the interview, felt very confused and disappointed in himself for not understanding gravity as well as he thought. Was very self-aware.
F12-10 Mechanical Engineering and Technology, 1 previous	- Misapplication of scientific model: Minimum mass requirement: Astronauts in Q2 do not feel a force because they are floating. Considers size and density in addition to mass and distance, reasoning that things that are larger and denser cause a stronger gravitational force (such as in Q11, Q10, Q14, and Q16). Inconsistent application inverse square law. Used it correctly for Q4, but then seemed to forget about it for the rest of the questions. Knew there was a force equation but couldn't remember it. - Mixing of Forces. Magnetism may be the cause of gravity (Q15), it helps hold us to Earth (Q8). - Directionality: On Q5 and Q12, reasoned that the force would always be perpendicular to the surface.
F12-11 Plant Biotechnology, 1 previous course	- Scientific Model: Correctly uses proportional reasoning with both mass and distance (Q4, Q7, Q10, Q16, Q25). - No Mixing of Forces Model: Only gravity holds people to Earth, atmosphere and magnetism have nothing to do with it, but expressed a little uncertainty about magnetism. - Understood that there is no minimum mass requirement (Q2) and no maximum distance (Q24). - Understood the directionality of gravity in all cases. Noted that in physics one always draws the weight force down on free body diagrams. - Misapplication of mass: at first reasoned that higher density means stronger gravity (Q14) because he thought it meant higher mass, but then decides that you can have different densities but the same mass (Q17), so changed his answer to Q14. - Wondered about the difference between gravitational force and gravitational acceleration. Decided that gravity was the same for all planets in Q3 because g = 9.8 m/s^2 for all objects on Earth. Thought that the mass of the satellite was needed for Q19.

APPENDIX T

IRT PARAMETER ESTIMATION

IRT software programs such as BILOG-MG estimate item and person parameters by maximizing the likelihood function of observed response patterns (Harris 1989). So, what is the likelihood function? First note that, with an assumption of local item independence, the probability function for a particular response pattern $\vec{u} = (x_1, x_2, \dots, x_N)$, where x_i can be either 1 or 0, given a student ability θ and a set of item parameters $\vec{a}$ and $\vec{b}$ is simply the product of the independent probabilities for responses to each question (Hambleton & Swaminathan 1985):

$$P(\vec{u}|\theta, \vec{a}, \vec{b}) = \prod_{i=1}^{N} [P_i(\theta)]^{x_i}[1 - P_i(\theta)]^{1-x_i}$$

In the four-group Marginal Maximum Likelihood (MML) (Hambleton & Swaminathan 1985) method that was used for the analysis of the NGCI, item parameters were estimated first by integrating out the θ dependence with an assumed distribution function for the pre-instruction population, $g(\theta)$, chosen to be Gaussian with a mean of zero and standard deviation of one (Bock & Lieberman 1970). This anchors ability estimates and provides a marginal probability function that depends only on item parameters:

$$\bar{P}(\vec{u}|\vec{a}, \vec{b}) = \int_{-\infty}^{\infty} P(\vec{u}|\theta, \vec{a}, \vec{b}) g(\theta) d\theta$$

There is no closed-form solution for this integral, and numerical integration must be performed (Bock & Lieberman 1970) by weighting the distribution function $g(\theta)$ at q nodes X_k with corresponding values $A(X_k)$:

$$\bar{P}(\vec{u}|\vec{a}, \vec{b}) = \bar{P}_u = \sum_{k=1}^{q} P(\vec{u}|X_k, \vec{a}, \vec{b}) A(X_k)$$

Recall that this function is only for one particular response pattern. To model the likelihood of all observed response patterns, one must multiply the marginal probabilities for the number of examinees r_u for each of the 2^N possible response patterns u:

$$L = \prod_{u=1}^{2^N} \overline{P_u}^{\,r_u}$$

Taking the logarithm leads to the more manageable marginal likelihood function:

$$lnL = \sum_{u=1}^{2^N} r_u \, ln\overline{P_u}$$

This equation can be differentiated with respect to the parameters a_i and b_i and maximized to obtain item parameter estimates (Zimowski et al 1996). The mathematical details of this maximization procedure can be found in Bock and Lieberman (1970) and Bock and Aitkin (1981).

Once item parameters have been estimated, they are taken as known quantities for procedures that estimate student abilities. NGCI student abilities were estimated using the expected *a posteriori* (EAP) method that is recommended by Harwell and Baker (1991) and follows the precedent of Wallace and Bailey (2010). The EAP method modifies the likelihood function of a response pattern (with now known parameters) by multiplying by the assumed population distribution (in this case, a Gaussian). It then maximizes this “posterior distribution,” which may not be a Gaussian, and assigns the mode of this distribution as the ability value for students who exhibit that unique response pattern. For the mathematical details of this procedure, one may refer to Baker and Kim (2004).

APPENDIX U

YEN'S Q3 STATISTIC FOR THE 2PL IRT ANALYSIS

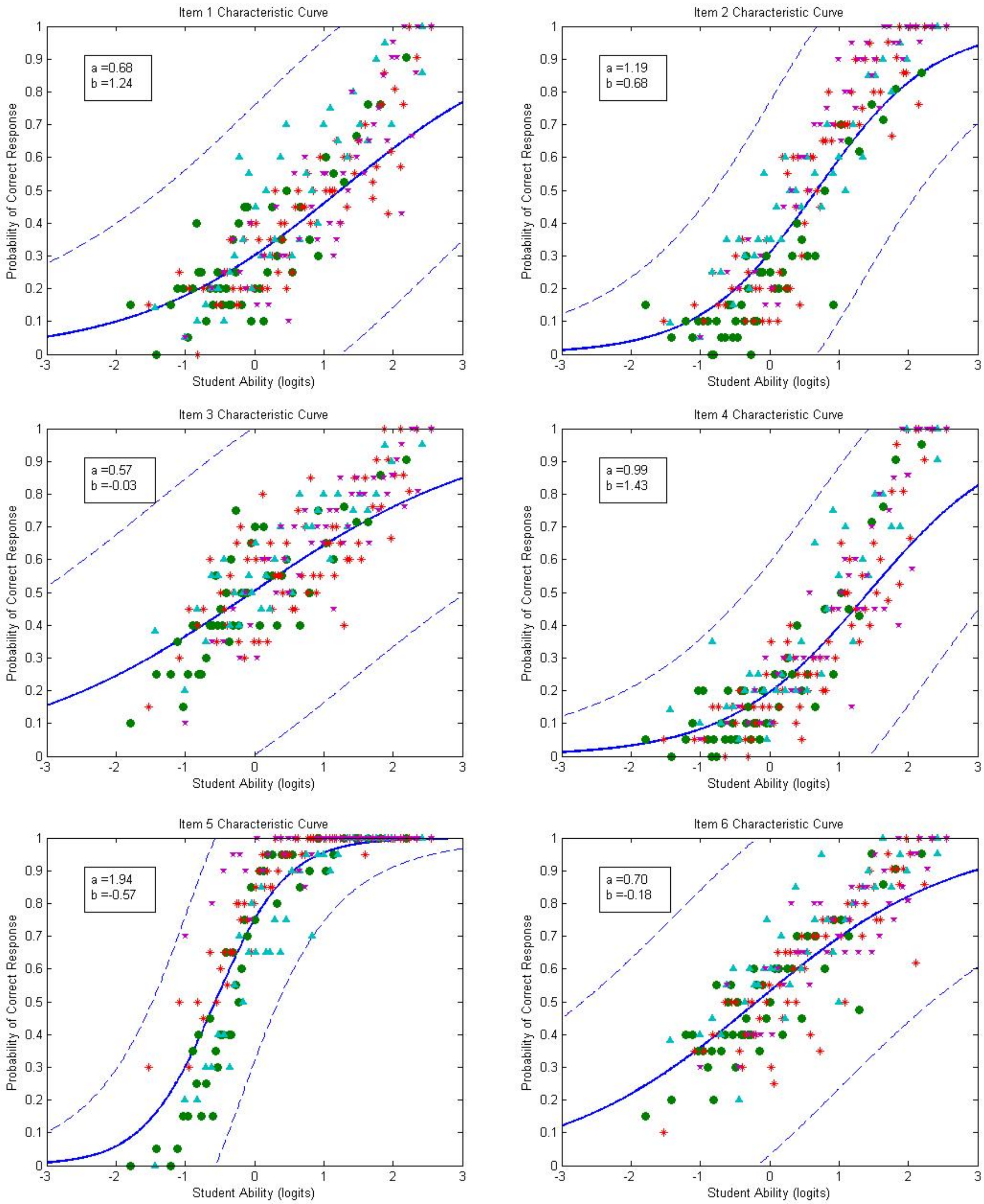
Yen’s Q3 values after eliminating Items 7, 10, 14, and 17 from the 2PL IRT analysis.

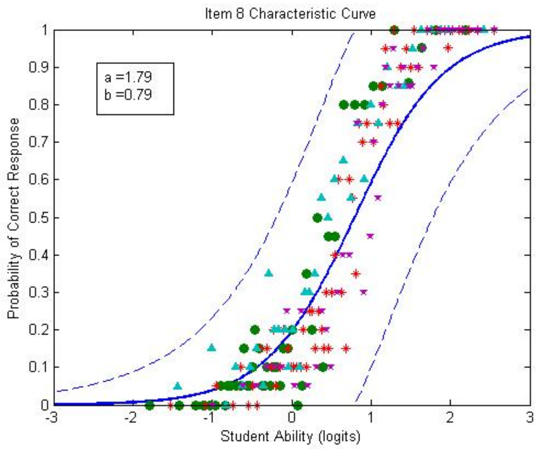
Item	1	2	3	4	5	6	8	9	11	12	13	15	16	18	19	20	21	22	23	24	25	26
1	1.00	-0.07	-0.04	-0.03	-0.03	-0.03	0.06	0.04	-0.01	-0.04	-0.04	0.10	0.01	-0.03	-0.07	-0.06	-0.08	-0.03	-0.07	-0.02	-0.02	-0.03
2		1.00	-0.01	-0.02	-0.08	0.01	-0.01	-0.09	-0.03	-0.05	-0.05	-0.09	-0.05	0.00	-0.06	-0.08	0.02	-0.09	-0.02	-0.04	0.01	-0.07
3			1.00	-0.03	-0.04	-0.02	-0.06	-0.02	-0.05	0.06	0.00	-0.09	0.00	0.15	0.04	-0.04	-0.04	-0.03	-0.07	-0.05	0.00	-0.04
4				1.00	-0.05	0.00	-0.05	-0.03	-0.05	-0.04	-0.01	-0.04	0.16	-0.01	-0.05	-0.07	-0.02	-0.08	0.00	-0.03	0.07	-0.07
5					1.00	-0.06	-0.11	-0.04	-0.04	-0.12	-0.04	-0.06	-0.04	-0.10	-0.10	-0.09	-0.10	-0.17	-0.08	-0.06	-0.03	0.05
6						1.00	-0.07	-0.02	-0.06	-0.01	-0.03	-0.04	-0.03	-0.04	-0.02	-0.03	0.00	-0.05	0.01	0.01	0.02	-0.07
8							1.00	-0.01	-0.03	-0.06	-0.05	0.09	-0.06	-0.07	-0.12	-0.15	-0.10	0.00	-0.13	-0.06	-0.05	-0.12
9								1.00	0.01	0.02	-0.01	0.01	0.00	0.00	-0.02	-0.05	-0.05	-0.06	-0.05	-0.03	-0.03	0.00
11									1.00	0.03	-0.04	-0.01	0.01	-0.03	-0.03	-0.02	-0.02	-0.03	-0.04	-0.02	-0.11	-0.02
12										1.00	-0.02	-0.06	-0.04	0.01	0.03	-0.03	-0.13	-0.01	-0.10	-0.06	-0.04	-0.03
13											1.00	-0.04	-0.04	0.02	0.00	-0.04	-0.03	-0.06	-0.03	-0.06	0.01	-0.01
15												1.00	-0.05	-0.08	-0.06	-0.07	-0.10	-0.03	-0.11	-0.01	-0.02	-0.05
16													1.00	-0.03	-0.06	-0.05	0.04	-0.02	-0.03	0.01	0.08	-0.04
18														1.00	0.05	-0.05	-0.02	-0.03	-0.02	-0.03	0.00	-0.03
19															1.00	0.05	-0.09	-0.06	-0.02	-0.05	-0.02	-0.06
20																1.00	-0.12	-0.06	-0.06	-0.01	0.00	-0.12
21																	1.00	-0.09	0.16	0.03	0.02	-0.04
22																		1.00	-0.11	-0.07	-0.04	-0.11
23																			1.00	0.08	-0.01	-0.01
24																				1.00	0.03	-0.02
25																					1.00	-0.03
26																						1.00

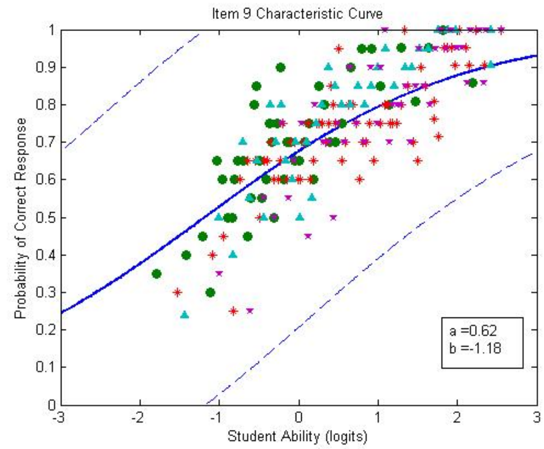
APPENDIX V

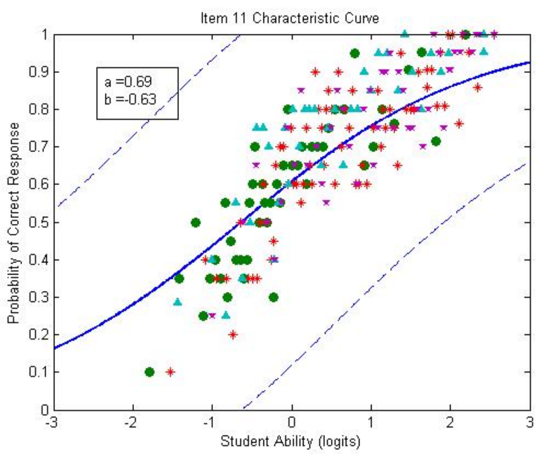
ITEM CHARACTERISTIC CURVES OVERLAID WITH GROUP DATA

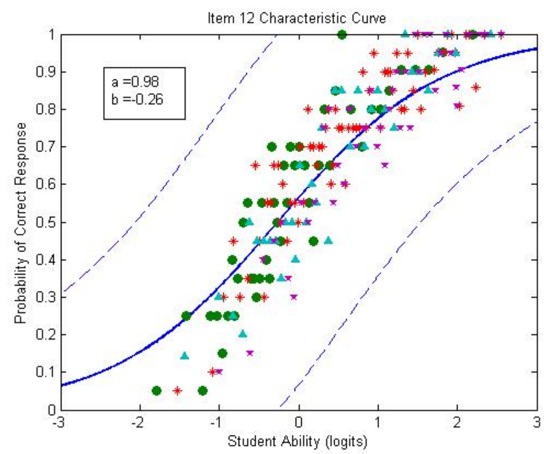
Green circles represent pre-instruction astronomy students, red asterisks represent pre-instruction physics students, turquoise triangles represent post-instruction astronomy students, and purple stars represent post-instruction physics students. Each data point represents approximately 20 students.

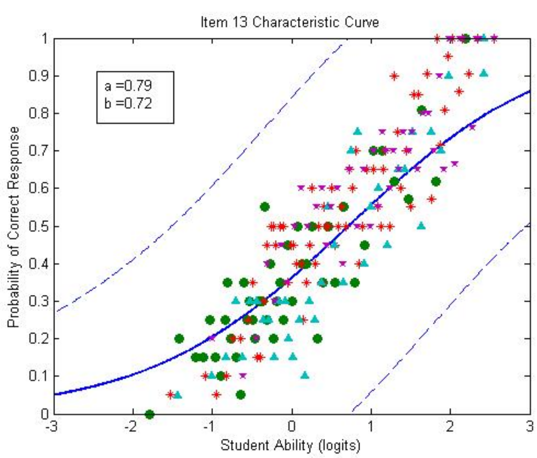


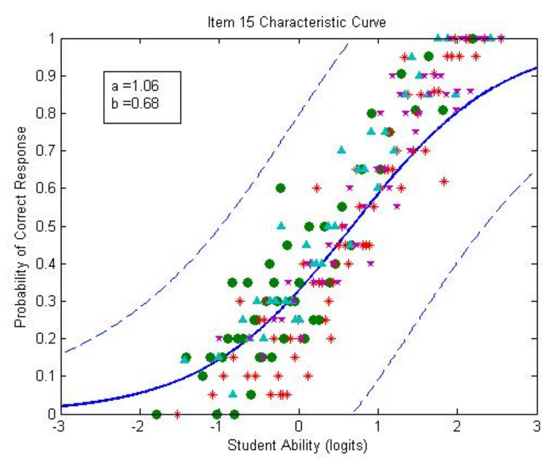

Item 8 Characteristic Curve
a =1.79
b =0.79
Probability of Correct Response
Student Ability (logits)

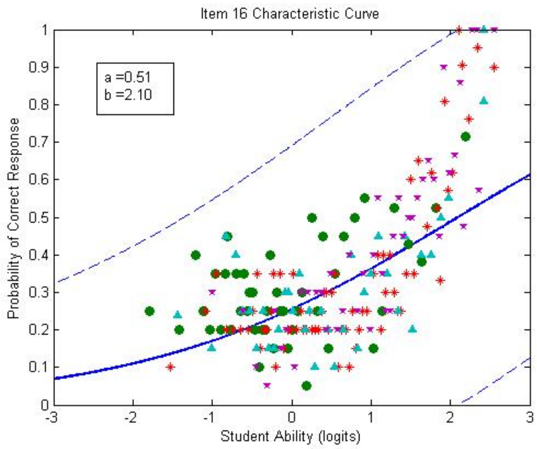

Item 9 Characteristic Curve
a =0.62
b =-1.18
Probability of Correct Response
Student Ability (logits)

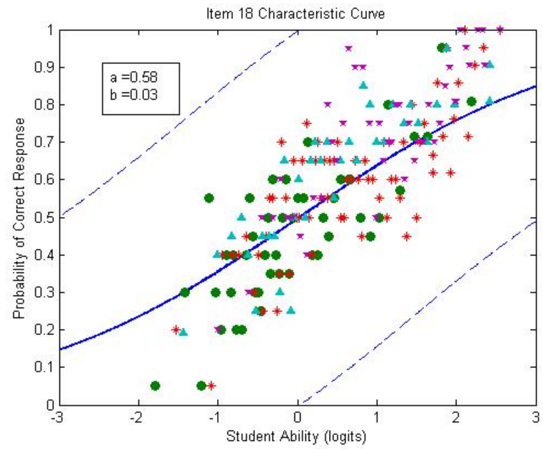

Item 11 Characteristic Curve
a =0.69
b =-0.63
Probability of Correct Response
Student Ability (logits)

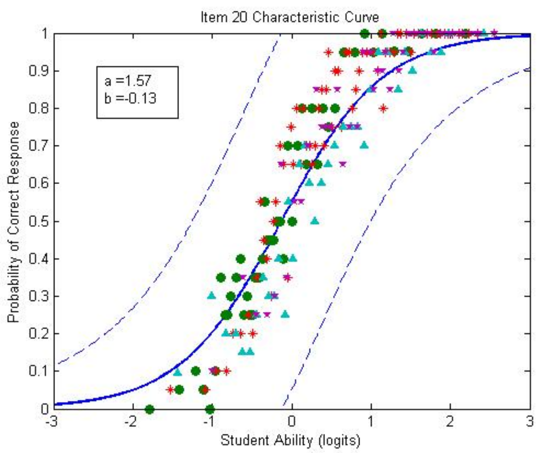

Item 12 Characteristic Curve
a =0.98
b =-0.26
Probability of Correct Response
Student Ability (logits)

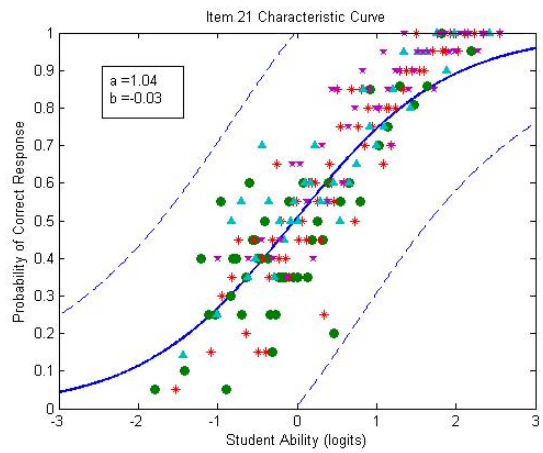

Item 13 Characteristic Curve
a =0.79
b =0.72
Probability of Correct Response
Student Ability (logits)

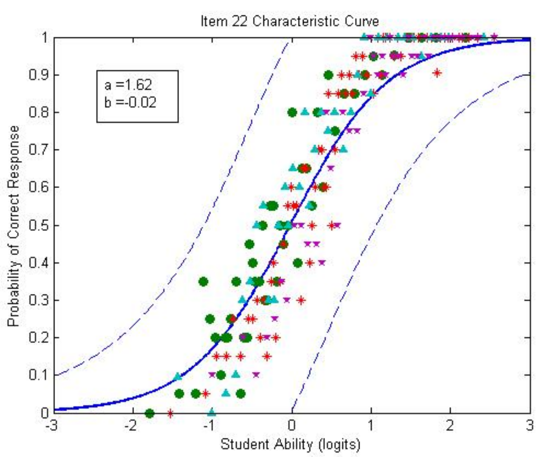

Item 15 Characteristic Curve
a =1.06
b =0.68
Probability of Correct Response
Student Ability (logits)

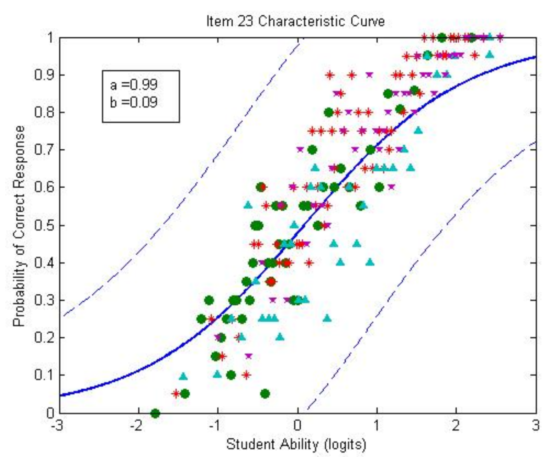

Item 16 Characteristic Curve
a =0.51
b =2.10
Probability of Correct Response
Student Ability (logits)

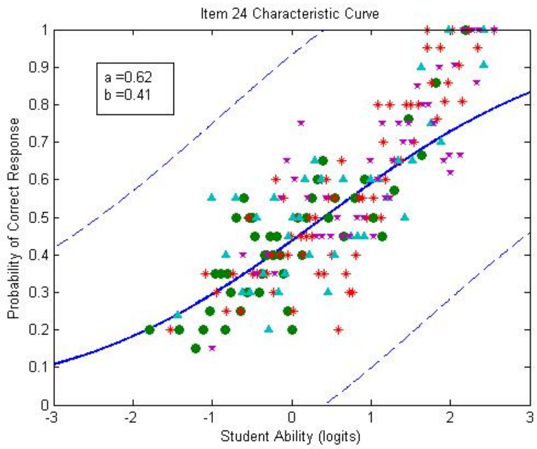

Item 18 Characteristic Curve
a =0.58
b =0.03
Probability of Correct Response
Student Ability (logits)

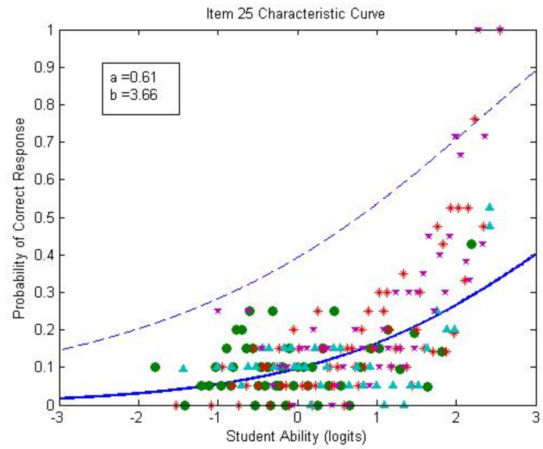

Item 20 Characteristic Curve
a =1.57
b =-0.13
Probability of Correct Response
Student Ability (logits)

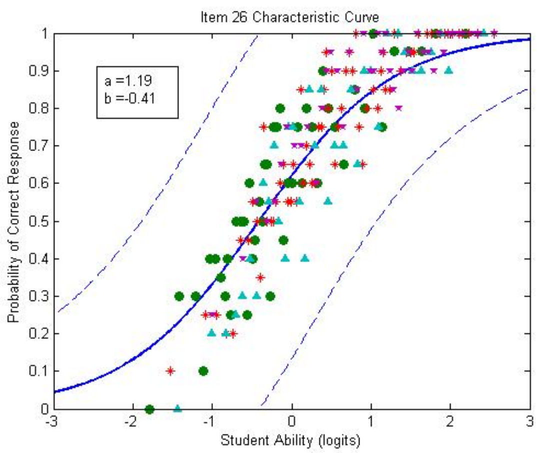

Item 21 Characteristic Curve
a =1.04
b =-0.03
Probability of Correct Response
Student Ability (logits)


Item 22 Characteristic Curve
a =1.62
b =-0.02
Probability of Correct Response
Student Ability (logits)


Item 23 Characteristic Curve
a =0.99
b =0.09
Probability of Correct Response
Student Ability (logits)


Item 24 Characteristic Curve
a =0.62
b =0.41
Probability of Correct Response
Student Ability (logits)


Item 25 Characteristic Curve
a =0.61
b =3.66
Probability of Correct Response
Student Ability (logits)


Item 26 Characteristic Curve
a =1.19
b =-0.41
Probability of Correct Response
Student Ability (logits)

APPENDIX W

NGCI INSTRUCTOR SURVEY

NGCI Instructor Survey

Please answer the following questions according to the semester that you used the NGCI in your class.

Date:
Name:
Number of Years Teaching:
Institution Name:
Type of Institution (Choose One):

a) 2 year college
b) 4 year college
c) Research university

1. Is your institution on a **quarter** or **semester** system?

2. How many total contact hours do you have **each week** with your class? (Do not include office hours.)

3. How many hours do you spend each week in the following instructional settings?
   a. Lecture-style setting, which may include questions and activities
   b. Recitation or discussion section
   c. Doing traditional laboratory investigations and astronomical observations
   d. Other (please specify)

4. How many hours during the **term** are used for **exams** (not counting the final exam)?

5. During the **term**, how many times do students work in groups on *collaborative learning* activities, excluding traditional labs (e.g., Tutorials in Introductory Physics, Astronomy Lecture-Tutorials, Ranking Tasks, Case Studies)? Provide a range if necessary.

6. During a **typical class**, how many times do you pose a question that requires your students to work together (e.g., Think-Pair-Share or Concept Test-type question)? Provide a range if necessary.

7. During a **typical class**, how many times are students asked to answer a question or make a prediction by themselves, without working together? Provide a range if necessary.

8. During a **typical class**, are there any other things that you do to promote student learning beyond lecture?
   a. Provide a description
   b. State how many times **in a term** this occurs