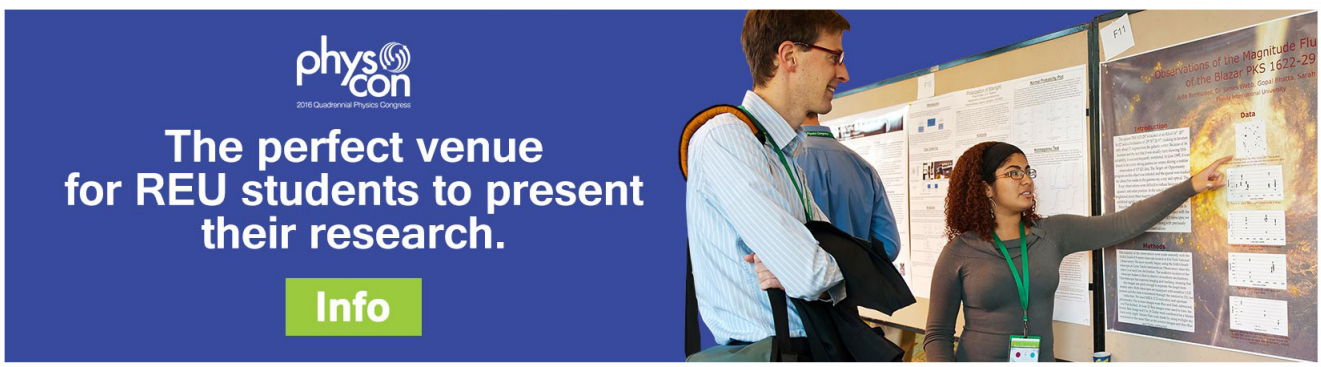
AAPT PHYSICS EDUCATION | AMERICAN JOURNAL of PHYSICS

Applicability of the Newtonian gravity concept inventory to introductory college physics classes

Kathryn Williamson, Edward E. Prather, and Shannon Willoughby

Citation: American Journal of Physics **84**, 458 (2016); doi: 10.1119/1.4945347
View online: http://dx.doi.org/10.1119/1.4945347
View Table of Contents: http://scitation.aip.org/content/aapt/journal/ajp/84/6?ver=pdfcov
Published by the American Association of Physics Teachers



PHYSICS EDUCATION RESEARCH SECTION

The Physics Education Research Section (PERS) publishes articles describing important results from the field of physics education research. Manuscripts should be submitted using the web-based system that can be accessed via the American Journal of Physics home page, http://ajp.dickinson.edu, and will be forwarded to the PERS editor for consideration.

Applicability of the Newtonian gravity concept inventory to introductory college physics classes

Kathryn Williamson
West Virginia University, Morgantown, West Virginia 26501

Edward E. Prather
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721

Shannon Willoughby
Montana State University, Bozeman, Montana 59717

(Received 15 April 2015; accepted 21 March 2016)

The study described here extends the applicability of the Newtonian Gravity Concept Inventory (NGCI) to college algebra-based physics classes, beyond the general education astronomy courses for which it was originally developed. The four conceptual domains probed by the NGCI (Directionality, Force Law, Independence of Other Forces, and Threshold) are well suited for investigating students' reasoning about gravity in both populations, making the NGCI a highly versatile instrument. Classical test theory statistical analysis with physics student responses pre-instruction (N = 1,392) and post-instruction (N = 929) from eight colleges and universities across the United States indicate that the NGCI is composed of items with appropriate difficulty and discrimination and is reliable for this population. Also, expert review and student interviews support the NGCI's validity for the physics population. Emergent similarities and differences in how physics students reason about gravity compared to astronomy students are discussed, as well as future directions for analyzing the instrument's item parameters across both populations. © *2016 American Association of Physics Teachers.*
[http://dx.doi.org/10.1119/1.4945347]

I. INTRODUCTION

Constructivism posits that learners incorporate new knowledge into existing mental landscapes,[1] and that in order to maximize learning, instructors need to be aware of their students' mental landscapes so that they can provide students with opportunities for "cognitive dissonance."[2] Students' existing mental landscapes can be described as "knowledge in pieces,"[3] organized as generative mental structures,[4] or the more-generalized "mental model" as "a robust and coherent knowledge element or strongly associated set of knowledge elements."[5] Taking a scholarly approach to studying students' discipline knowledge development, researchers in physics education and astronomy education have spent considerable time developing reliable and valid assessments within critical disciplinary topics. Pre- and post-instructional testing with concept inventories have become a common method for assessing physics and astronomy teaching and learning on a large scale, leading to profound insights and evidence for the success of curriculum reform.[6–13] In this paper, we build on this literature and provide a new tool for assessing changes in introductory algebra-based students' conceptual understandings and reasoning about gravity—the Newtonian Gravity Concept Inventory (NGCI).

The NGCI was iteratively developed with a broad population of general education introductory astronomy (hereafter "astronomy") students, but its conceptual focus is well aligned with the introductory algebra-based physics (hereafter simply "physics") curriculum. While it may be taught explicitly in only one or two lectures, the concept of gravity is applied widely in treatments of projectile motions and weight forces, and within the context of Newton's Laws, conservation of energy, and conservation of momentum. Physics instructors and researchers can gain some insights about their students' understanding of gravity by assessing students' understanding of forces using existing assessments such as the Force Concept Inventory,[14] but these tend to focus on the effects of gravity on or near Earth's surface. Our research on astronomy students has illustrated that there are unique reasoning difficulties related to thinking about gravity that are beyond difficulties related to forces in general,[15] however, research on the prevalence of these gravity-specific reasoning difficulties within the physics population has not been previously conducted. This paper investigates the applicability of the NGCI as an appropriate and useful tool for assessing the conceptual and reasoning difficulties with physics students.

The physics population represents a vastly different group than the astronomy population in terms of interest, motivation, comfort with math and science, and career path. Most physics students are science majors and have had previous instruction in astronomy or physics, whereas astronomy

students are primarily non-science majors who take the course as a one-time general education science requirement before going on to become society's journalists, politicians, lawyers, historians, business leaders, and teachers.[11,16] Also, the typical physics course investigates Newtonian gravity within a very different context than the astronomy course. (The astronomy curriculum emphasizes gravity in the context of planet and star formation, space travel, the structure of galaxies, and the evolution of the universe.) However, much of the literature on student understanding of gravity has actually focused on physics students,[17–19] so investigating the applicability of the NGCI to the physics population is a natural extension.

Care must be taken in extending the applicability of an assessment instrument from one population to another, because reliability and validity measures are inextricably set in the context of a particular population. Differences in students' backgrounds, motivations, interests, or future career aspirations may significantly affect their responses in unpredictable or conflicting ways. If instructors and researchers uncritically use a concept inventory without understanding its applicability, they may make false inferences about their students' learning and the effectiveness of instructional interventions. Therefore, before using the NGCI to assess physics students' understanding of gravity, we must conduct a carefully designed research study using the instrument with this population and ask: How internally consistent is the NGCI for the physics population? Are NGCI items of appropriate difficulty and discriminatory power for the physics population? How do distractor choices function? How similar or different are physics students' response patterns to those of astronomy students? Answering these questions will provide instructors of introductory algebra-based physics with an understanding of how to use the NGCI to assess their students' conceptual and reasoning difficulties related to gravity. The study described here also can serve as an example for the discipline-based education research community on the methods for testing the applicability of assessment instruments across populations.

The Background Section of this paper recaps the development process of the NGCI with astronomy students.[15,20] We discuss the conceptual focus of the NGCI along four domains (Directionality, Force Law, Independence of Other Forces, and Threshold) and argue for the applicability of these domains across both the astronomy and physics curricula. Next, in the Method Section, we establish the applicability of the NGCI to physics courses through testing the NGCI with a large sample of physics students at eight colleges and universities across the United States, providing a dataset of 1,392 pre-instruction responses and 929 post-instruction responses. In the Results Section, we discuss our classical test theory analysis and argue that the NGCI is a robustly reliable instrument for the algebra-physics population, with items of appropriate difficulty and useful discrimination capabilities. In the Validation Section, we draw on student interviews and expert analysis, including that done during the initial development with astronomy students,[20] to argue for the validity of the NGCI in assessing physics students' understanding of gravity. The Discussion Section investigates physics students' performance on the NGCI in comparison to astronomy students' performance. Emergent trends in the similarities and differences with how physics and astronomy students reason about gravity are discussed within each of the four conceptual domains. Finally, we conclude with implications for instruction, emphasizing the unique capabilities of the NGCI as a cross-disciplinary instrument and projecting future research directions.

II. BACKGROUND

Our previous publications[15,20] discussed the development of the NGCI for astronomy courses following the four-phase model of instrument development by Benson and Clark:[21] Planning, Construction, Quantitative Evaluation, and Validation. The grounded theory[22,23] phenomenographic coding of student responses to open-ended questions described in Ref. 15 identified prevalent alternative mental models of gravity (a Boundary Model, an Orbital Indicator Model, and a Mixing of Forces Model), as well as misapplications of the scientific model (for example, confusions about the measurement of distance and the difference between mass and density). The reader is encouraged to consult Ref. 15 for full descriptions of these conceptions. The conceptions then informed phrasing and distractor choices for multiple-choice items. Survey items were written following the best practices for item construction outlined in Haladyna *et al.*,[24] piloted iteratively with astronomy students, and evaluated using a set of commonly accepted methods for evaluation of multiple-choice conceptual surveys.[6,11–13,25] Three survey iterations resulted in the full 26-item Newtonian Gravity Concept Inventory (NGCI). Classical Test Theory statistics show that the NGCI has high reliability (Cronbach's Alpha = 0.84) and has appropriate item difficulty and discrimination. Expert agreement of item functioning (Fleiss's Kappa = 0.80), a high average score of experts (96.7%), and "think-aloud" interviews with twenty-four astronomy students were among the evidence used to argue for the validity of the NGCI.[26] Again, we encourage readers interested in learning more about the rigorous qualitative and quantitative methods used to create and evaluate a concept inventory of this type to consult Refs. 15 and 20, and references therein.

The conceptual focus of the NGCI is captured in four primary domains: (1) The Directionality Domain, (2) the Force Law Domain, (3) the Independence of Other Forces Domain, and (4) the Threshold Domain. The Directionality Domain probes understanding of the direction of gravitational force in situations with multiple objects, relative motion, and for objects on the surface of a large body. NGCI items probing the Directionality Domain investigate whether the direction of the gravitational force should be determined by superposition, the direction of apparent weight, and if the gravitational force is always perpendicular to the surface. The Force Law domain investigates students' reasoning about how the magnitude of gravitational force is determined, including the role of distance and mass, and the effects of changes in density (for example, whether distance is measured from the center of mass of an object, from the surface of an object, or if it is simply the radius of the object). The Independence of Other Forces Domain probes the well-documented naïve idea that gravity is confounded with forces associated with air pressure, magnetism, and rotation.[14,27–30] The Threshold Domain probes student understanding of the universality of gravity, particularly in limiting cases such as large distances or small masses. This domain also investigates whether students believe there to be a "boundary" for which gravity suddenly changes or stops, such as the "edge" of the atmosphere or an orbital path. Other work[20] provides much greater detail into the breadth of both correct and incorrect ideas related to these

Table I. Physics demographic data for the NGCI, calculated as averages from 1,392 pre-instruction responses and 929 post-instruction responses (Ref. 33).

		%
Major	Business	0.8
	Education	2.0
	Humanities, Social Sciences, Arts	3.2
	Science, Engineering, Architecture	87.4
	Other	6.5
Previous Physics or Astronomy 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

four conceptual domains, as well as insight into how different student ideas are elicited by the contexts of specific questions.

III. METHOD

To investigate the reliability and validity of the NGCI for the physics population, the instrument must undergo pilot testing, quantitative analysis, and validation, just as was necessary when it was first developed for the astronomy population. To this end, 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 (all four-year institutions). In total, 1,392 Physics students participated in the NGCI pre-instruction, and 929 Physics students participated in the post-instruction (Table I shows the demographic data). As with the astronomy data in Ref. 20, demographics percentages were calculated as averages from both the pre- and post-instruction responses 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. Demographic data are consistent with other estimates,[31] so this sample is likely representative of physics students nationwide.[32]

Table II provides NGCI descriptive statistics, class-averaged normalized gains, and class effect sizes where appropriate, for these physics pilot sites. Average class pre-instruction scores ranged from 43% to 71%, with a total population average score of 58%, and average class post-instruction scores ranged from 50% to 85%, with a total population average score of 68%. Class-averaged normalized gains ranged from 0.12 to 0.52, and class effect sizes ranged from 0.38 to 0.98. A short survey given to the instructors indicates that the level of student-centered interactivity in the class supports the observed gains (for example, UC Santa Barbara had the highest level of interactivity, likely explaining the higher learning gains). As with the astronomy pilot testing data, these broad ranges of pre and post scores, normalized gains, and effect sizes in Table II serve as a "first order" indication that the NGCI is sensitive to a wide range of physics students' understanding of Newtonian gravity and differences in physics instruction.

NGCI reliability and item-functioning with the physics population are assessed with Classical Test Theory (CTT) statistics. CTT provides measures of internal consistency, item difficulty, and item discrimination. While these measures were calculated during the development of the NGCI from a national sample of astronomy students,[20] CTT is highly sample dependent[34,35] and must be re-calculated for the physics sample. Therefore, the NGCI has a different reliability index and different item difficulty and discrimination parameters for physics students than it does for astronomy students. Reliability is measured by the Cronbach's Alpha internal consistency statistic. Values range from 0 to 1 and are highest when the variance of total test scores is large compared to the variance within each item (i.e., the survey assesses differences across students rather than differences across items). Values over 0.70 are conventionally accepted as being internally reliable.[25,36] Item difficulty is calculated as the percentage of students who answered *incorrectly*, such

Table II. 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 University	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 Pre	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

that a very high percentage indicates that most students answered incorrectly (i.e., the item may be too difficult) and a very low percentage indicates that most students answered correctly (i.e., the item may be too easy). In order for a survey to be of appropriate difficulty, most items should have difficulty values between 0.20 and 0.80.[13,37] Finally, item discrimination is calculated with the point-biserial correlation between student performance on the item and their overall performance on the NGCI.[25,34] For item discrimination, anything greater than 0.30 indicates that students' scores on that item are well-correlated with their total scores.[13,25]

Additional information from student response patterns and interviews is critical for providing context to these statistics and to further the discussion of the validity of the NGCI. To this end, we provide histograms of the proportions of students choosing each answer choice for each of the 26 NGCI items in Supplement A (available online[38]), as well as interview notes from eleven "think-aloud" interviews with physics students[39] in Supplement B (also online[38]). As with the astronomy student interviews from Ref. 20, these physics student interviews followed the protocols and methods of Bolton and Bronkhorst,[40] in which the interviewer engaged in "back channeling." This technique prompted students to continue talking out loud to elaborate or clarify their reasoning while working through the NGCI. Notes were taken during interviews and additional thoughts were recorded as soon as possible after the interview ended. We unpack the meaning of these qualitative results in context with the CTT statistics in the Discussion Section.

IV. RESULTS

For the physics pilot data, the NGCI's pre-instruction Cronbach's Alpha reliability is 0.82 and the post-instruction reliability is 0.86. These Cronbach's Alpha values are very high (actually higher than the Astronomy values of 0.79 and 0.84, respectively,[20] indicating that the NGCI is a reliable instrument for the physics population; i.e., the instrument is sensitive to differences across students rather than differences across survey items. Theoretically, then, if a student took the test over many administrations their scores would be reliably similar. The CTT item difficulty and discrimination statistics for the physics sample are shown in Table III, with items having difficulty and discrimination values outside the conventionally accepted range bolded and italicized.

Table III. NGCI CTT item statistics for physics pilot site data. Item difficulty and item discrimination are shown 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.

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

From Table III, one can see that most items on the NGCI are of appropriate difficulty and discriminatory power for physics students, with some notable exceptions. Pre-instruction, one item (item 25) appears too difficult and one (item 5) appears too easy. Post-instruction, however, no items appear to be too difficult, but four items (items 5, 19, 20, and 26) appear to be too easy. Moreover, because 97% of physics students answer item 5 correctly post-instruction, this item is not a good discriminator of student understanding and reasoning ability.However, as we will discuss in Sec. V, this item serves as a good control question to ensure that students have moved forward as the result of instruction and are answering questions earnestly. All other items have good CTT discrimination values for the physics sample both pre and post-instruction.

What about the validity of using the NGCI for assessing physics students' understanding of gravity? While the NGCI was not developed with information from physics students' responses, we apply many of the same arguments for validity[20] that support the NGCI as a valid tool for measuring astronomy students' understanding of Newtonian gravity. First, the construct of Newtonian gravity, as captured in the four conceptual domains of the NGCI, is taught in both astronomy and physics courses. Second, because the development of the NGCI was strongly informed by prior research reported in the literature with insight from both physics and astronomy experts, and because it was a group of physics experts who evaluated the correctness of questions and answers, we argue that the NGCI is an appropriate instrument for probing physics students' understanding of gravity. Third, drawing heavily on the student response patterns to individual items (online Supplement A[38]) and student interviews (online Supplement B[38]), we can see that distractor choices function as intended; namely, as probes of known naïve ideas and reasoning difficulties. Indeed, Supplement A shows overall very similar functioning of distractor choices for both physics and astronomy students, and Supplement B links the specific mental models discussed in the Background (and in Ref. 15) to interviewee's individual thought patterns.

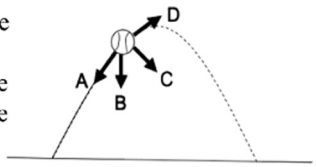
Taken together, the CTT reliability and item analysis, and the validity arguments in this section provide compelling evidence that the NGCI can provide meaningful assessment of physics students understanding of Newtonian gravity. Section V uses this data to show in finer detail how the NGCI is capable of providing valuable, nuanced information about the similarities and differences in how physics students reason about gravity compared to astronomy students.

V. DISCUSSION

In this section we offer some first findings of physics students' reasoning about gravity as assessed by the NGCI. In particular, we use the astronomy students' responses as a lens through which to understand the performance of physics students. We ask: How do physics pilot site data in this study compare to the astronomy data?[20] How do the CTT values compare? In context with the astronomy data, what do physics students' response patterns imply about their conceptual understanding and reasoning about gravity?

By comparing the demographics, descriptive statistics, and CTT parameters of the physics pilot sites (Tables I and II) to those of the astronomy pilot sites (Ref. 20,

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?



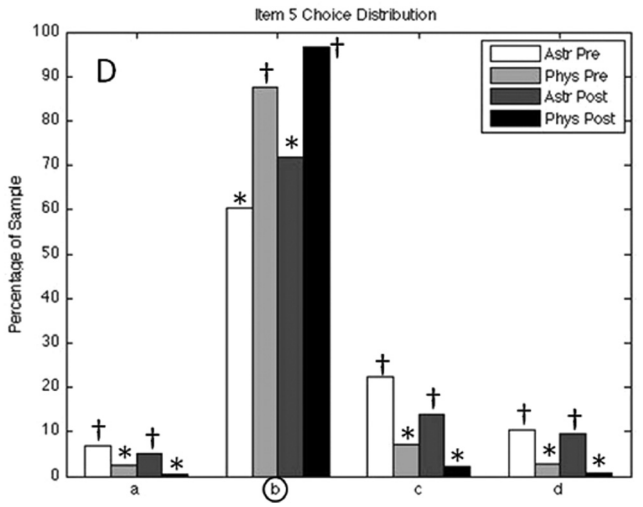
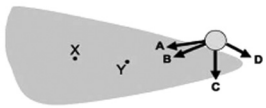


Fig. 1. NGCI item 5 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

Tables 2 and 3, and Figs. 1 and 2), one can notice interesting similarities and differences. First, the demographics of the samples reflect the differences in the types of students who take these courses that one might expect: 67.4% *more* of the physics sample reported a major of Science, Engineering, or Architecture, and 17.4% *more* of the physics sample had taken at least one previous astronomy or physics course. As a result of these differences, one would expect physics students to have a greater understanding of gravity. Indeed, average class scores are generally higher for the physics sample. The physics population-averaged pre-instruction and post-instruction scores were 14.17% and 12.47% *higher* than those of the astronomy population, respectively. Additionally, in comparing the CTT statistics, we see that the mean pre- and post-instruction item difficulty values calculated with the physics population are 0.14 and 0.13 *lower* than those for the astronomy sample, corroborating the finding that the NGCI is easier for the physics student population. However, the mean pre- and post-instruction physics item discrimination values were only 0.06 and 0.02 higher than those for the astronomy sample, implying that, overall, NGCI items function equally well in discriminating between students of low and high ability for both astronomy and physics students. These pieces of evidence show the versatility and robustness of the NGCI, and they provide a first indication that the students in introductory algebra-based physics courses have a greater overall understanding of gravity compared to students in general education astronomy courses.

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.



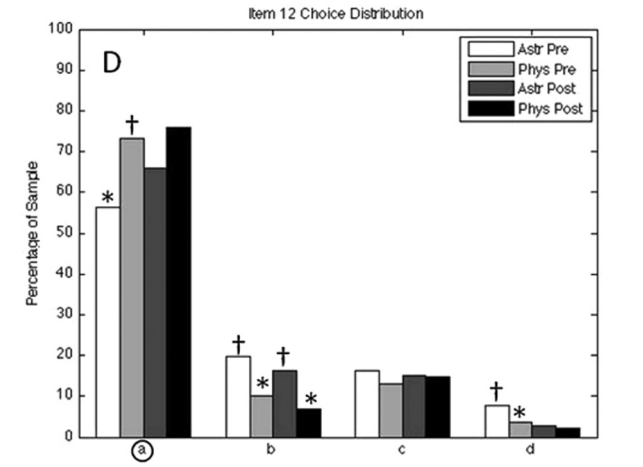


Fig. 2. NGCI item 12 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

However, the NGCI can provide information about the prevalence of known difficulties and naïve reasoning patterns of physics students beyond their overall stronger understanding of gravity. Supplement A shows response frequencies for each answer option to each question for the pre- and post-instruction physics data compared to the astronomy data from Ref. 20. Each item is labeled with its corresponding domain(s) (D = Directionality, FL = Force Law, OF = Independence of Other Forces, and T = Threshold) according to the expert categorization discussed in Ref. 20. On first inspection, the most noteworthy finding is how similar the response patterns are, implying that, despite the differences in the types of students who take physics compared to astronomy, and in addition to the vastly different contextual focus on Newtonian gravity in the introductory physics and astronomy courses, physics students are drawn in similar proportions as astronomy students to distractors that represent common naïve ideas and reasoning patterns. However, nuanced differences in how physics and astronomy students answer NGCI questions can be observed through chi squared statistical analysis with *post hoc* analysis.[41,42] Supplement A response proportions are marked as significantly (alpha < 0.05) lower than expected (asterisks) or higher than expected (daggers). Some items (such as items 9, 11, and 16) show few significant differences between physics and astronomy students' response patterns, whereas other items (such as items 2, 5, 8, 13, 21, 23, and 26) show that physics students are drawn to the correct answer more often than expected and drawn to the distractors less often than expected (and vice versa for astronomy students). In Subsections V A–V D, we offer some interpretive arguments about what these observed similarities and differences might imply about physics students' reasoning within each of the four NGCI conceptual domains. We emphasize that these interpretive arguments are not strong conclusions, but rather identify interesting differences that warrant further investigation. These arguments serve two purposes: (1) to expand our arguments above in favor of the validity and robustness of the NGCI for both physics and astronomy populations, and (2) to alert physics instructors and researchers to potential trends in their students' reasoning about gravity.

A. Directionality domain

Physics students appear to have a slight bias in reasoning that the direction of the gravitational force is "perpendicular to the surface." This is likely because these students have much greater familiarity with problems where gravity can be assumed to be "straight down." Item 5 (Fig. 1) and item 12 (Fig. 2) illustrate this best. As a projectile motion problem, item 5 is extremely easy for physics students, with a much higher proportion than expected answering correctly (97%). While one might conclude that this high proportion represents a significantly greater understanding in the Directionality Domain, response patterns about the direction of gravitational force on the non-spherical asteroid in item 12 indicate no significant difference in the proportions of physics and astronomy students choosing the correct direction for the gravitational force. Physics students chose "c" (beneath the ball) more often than "b" (toward the geometric center), whereas this trend was reversed for astronomy students. Interviews 4, 10, and 11 in Supplement B[38] indicate that this is likely associated with the idea that the direction of the gravitational force is "perpendicular to the surface," whereas this idea never came up in the astronomy student interviews.

B. Force law domain

Compared to astronomy students, physics students appear to have a greater propensity to reason quantitatively with mass and distance in determining force. Items 7, 10, and 16 show relatively equal difficulty values for physics and astronomy students, but these items have higher discrimination values when calculated with the physics sample. This implies that these items point to real reasoning differences. Item 10 (Fig. 3), for example, shows that, when reasoning about superposition, physics students are drawn preferentially to the correct answer ("d") and less distracted by the extreme choices ("a" and "e"). Astronomy students, however, show

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.

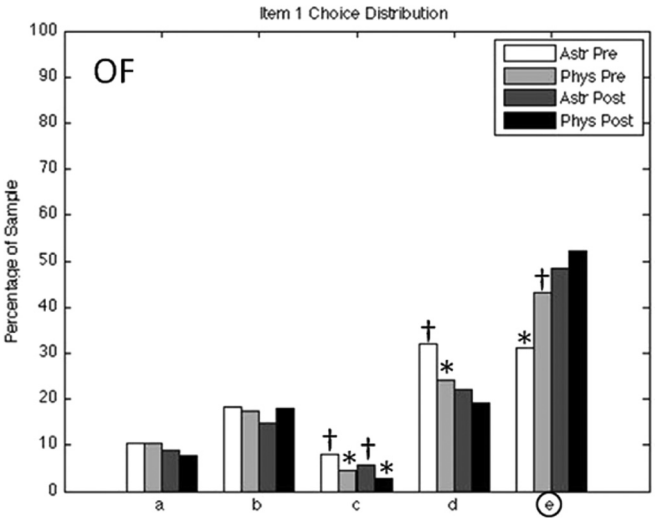
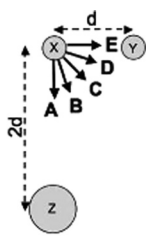


Fig. 4. NGCI item 1with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

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?



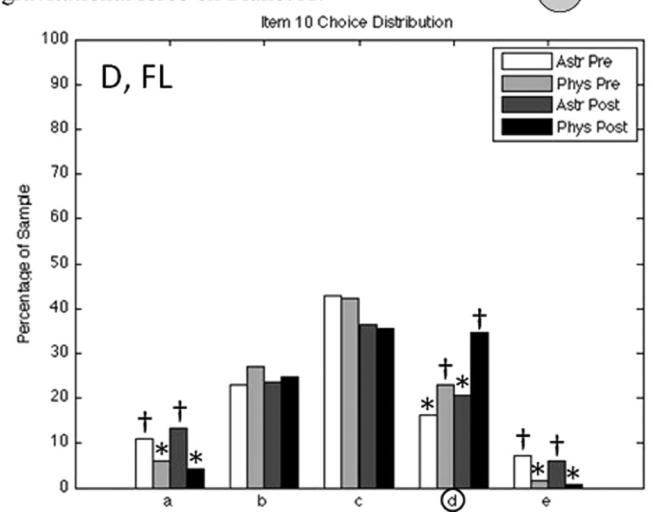


Fig. 3. NGCI item 10 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

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.

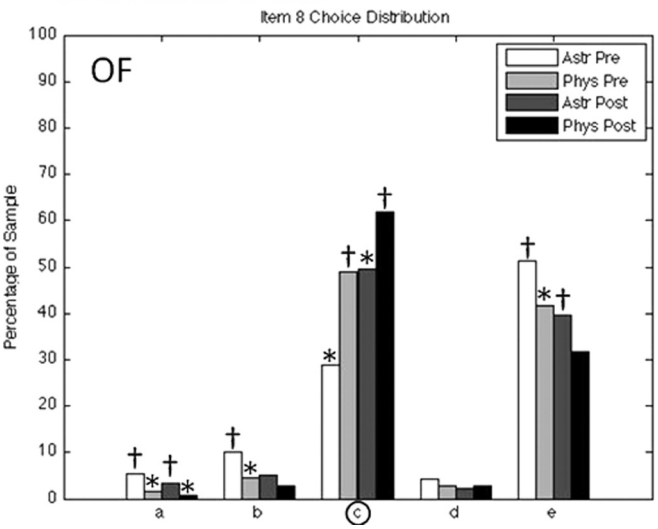


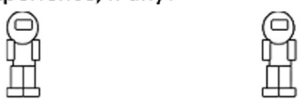
Fig. 5. NGCI item 8 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

more distribution in their choice of answers, with choice "c" chosen most often, indicating their preference for a linear relationship between gravitational force and both mass *and* distance. The larger than expected proportions of astronomy students choosing "a" and "e" indicate reasoning that an object can experience a gravitational force from only one of the other objects, which is either the *larger* object or the *closer* object, respectively. These more intuitive, non-quantitative choices are significantly less distracting to physics students.

C. Independence of other forces domain

Physics students appear to hold the well-documented naïve association between gravitational force and other factors including magnetism, rotation, and presence of an atmosphere just as strongly as astronomy students. Item 1 (Fig. 4) and item 8 (Fig. 5) best illustrate this point. Item 1 response patterns show no significant post-instruction differences in the proportions of physics and astronomy students choosing distractors except for the magnetism distractor ("c"). However, item 8 shows no significant post-instruction difference for the magnetism distractor ("b"), but it does for the atmosphere distractor ("a"). So, while the content of items 1 and 8 are almost identical, the response patterns show slightly different trends, indicating no consistent difference in how physics and astronomy students reason about gravity in context with other forces.

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.

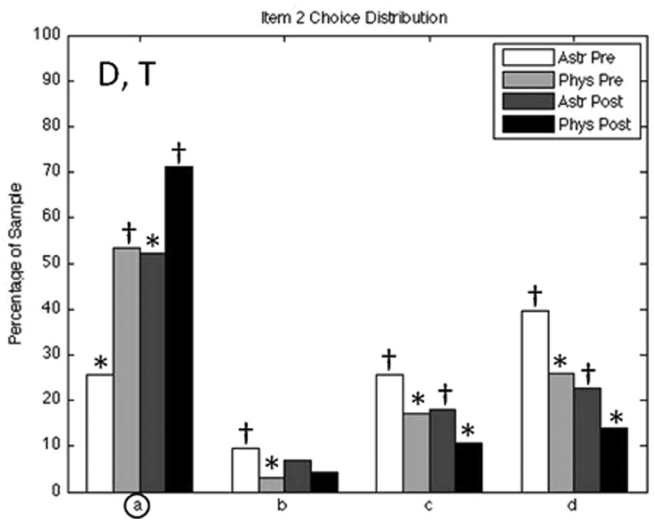
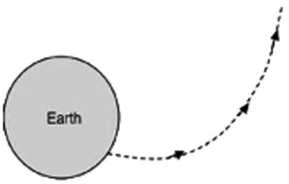


Fig. 6. NGCI item 2 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

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.

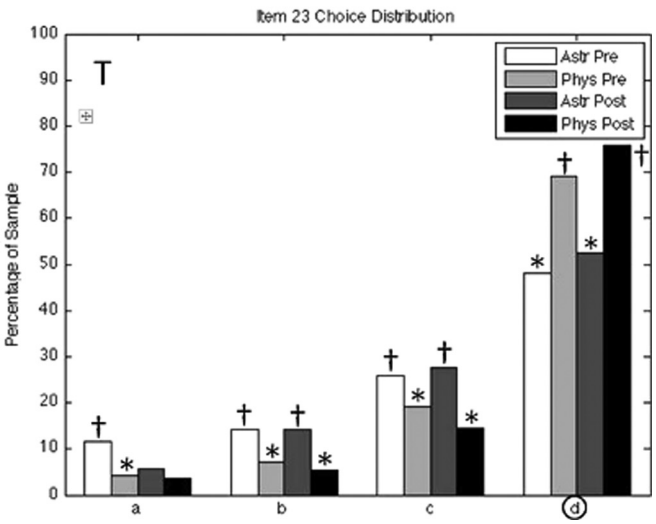


Fig. 7. NGCI item 23 with response patterns. Proportions marked with an asterisk are significantly lower than expected, and those with a dagger are significantly higher than expected.

D. Threshold domain

The Threshold Domain is where physics students excel most above astronomy students in understanding Newtonian gravity. After item 5 (discussed above in the Directionality domain), the next five most differentially functioning items on the NGCI are within the threshold domain (items 2, 13, 20, 21, and 23). With similar discrimination values, the difficulty values for these items were much lower for the physics sample. Item 2 (Fig. 6) and item 23 (Fig. 7) exemplify how physics students chose distractors that represent alternative ideas regarding a mass threshold (i.e., only "heavy" objects exert a gravitational force) or distance threshold (i.e., objects must be near each other to exert a gravitational force on one another) significantly less than expected. From response patterns in Supplement A[38] and student interviews, it appears that physics students have a more sophisticated understanding of the universality of gravity in limiting cases.

VI. CONCLUSIONS

This paper provides a robust investigation that establishes the applicability of the Newtonian Gravity Concept Inventory (NGCI) as a reliable, valid tool for assessing

student understanding of gravity beyond the introductory college astronomy population to the college algebra-based physics population. In the Background Section, we describe the development of the NGCI through Benson and Clark's[21] four-phase model using data from astronomy student responses. The Planning phase from Ref. 15 is discussed, as is the Development, Quantitative Evaluation, and Validation from Ref. 20. The Method Section of this paper details how we conducted a pilot study using the NGCI with a large sample of physics students from around the United States to check the instrument's applicability to this new student population. The Results Section shows the Classical Test Theory statistics calculated with physics student responses. The high pre- and post-instruction Cronbach's alpha shows that the NGCI is reliable, and item difficulty and discrimination indices show that most items are of appropriate difficulty and discrimination for this population of students. Supplemented by the strong evidence from student response patterns in Supplement A and interviews in Supplement B,[38] we assert that the NGCI functions as a valid assessment of physics student understanding of gravity.

The Discussion Section illustrates that, despite an overall stronger understanding of gravity, physics students overall display many of the same types of reasoning difficulties as astronomy students, with preliminary, nuanced patterns observed in each of the NGCI's four conceptual domains. In particular, physics students show a preference for a direction of the gravitational force that is "perpendicular to the surface." Physics students appear to implement some form of quantitative reasoning more frequently than astronomy students when determining the magnitude of the gravitational force. Additionally, physics and astronomy students seem to be equally drawn to distractors that represent a confounding of gravity with other forces. Finally, when reasoning about gravity, physics students excel most above astronomy students in scenarios that probe the universality of gravity in limiting cases. Further research may be needed to understand the effect sizes of these trends; however, this study illustrates that all four of the conceptual domains of the NGCI are uniquely sensitive to how student reasoning difficulties manifest for both astronomy and physics students. This serves to underscore the NGCI's robust validity across a diverse range of learners.

VII. FUTURE DIRECTIONS

As the only assessment instrument shown through robust research to be effective at assessing student understanding in both astronomy and physics courses, the NGCI has a unique versatility and is poised to inform curriculum reform in both disciplines. Student responses to the NGCI show the intuitive ways students reason about gravity that may or may not be consistent with the Newtonian perspective, allowing instructors to track how student understanding changes pre- to post-instruction. Furthermore, because gravity is such a foundational topic for a broad range of physical sciences, the NGCI could potentially be used beyond introductory astronomy and algebra-based physics. For example, how do students in calculus-based physics perform on the NGCI? What about more advanced physics classes? Gravity is a topic taught throughout the college physics curriculum, from intro- to upper-level courses; how does physics and astronomy majors' understanding of gravity progress over time? What are the best teaching practices to help learners develop a greater understanding of gravity? We invite practitioners and education researchers to help answer these questions by using the NGCI in their classes and following the methods here for adapting instruments to new populations. Copies of the NGCI can be requested by emailing the authors.

Finally, while Classical Test Theory statistics are sample-dependent, an Item Responses Theory (IRT) analysis allows for population-independent item parameters and item-independent student ability estimates. In future publications, we plan to combine the astronomy and physics students' response data in order to implement an IRT analysis.[43] Using the IRT item parameters, we will calibrate the instrument for all astronomy and physics students, and we will calculate student ability estimates so that we can compare students along a linear continuum. We then plan to implement a regression analysis on IRT student ability, controlling for student demographic data and course structure, to robustly tease apart which factors lead to the observed similarities and differences between astronomy and physics students' understanding of Newtonian gravity. This IRT analysis will also help us better understand which types of instructional methods lead to the highest learning gains for both populations of students.

ACKNOWLEDGMENTS

The authors would like to thank the physics instructors and their students who provided data for this research, as well as the reviewers for providing thoughtful feedback that significantly improved the quality of this manuscript.

[1]*Constructivism: Theory, Perspectives, and Practice*, edited by C. T. Fosnot (Teachers College Press, New York, 1996).
[2]*How People Learn: Brain, Mind, Experience, and School*, edited by J. D. Bransford, A. L. Brown, and R. R. Cocking (National Academy of Sciences, Washington, DC, 1995).
[3]A. diSessa, "Toward an epistemology of physics," Cognit. Instr. **10**, 105–225 (1993).
[4]S. Vosniadou, "Capturing and modeling the process of conceptual change," Learn. Instr. **4**, 45–69 (1994).
[5]L. Bao and E. F. Redish, "Model analysis: Representing and assessing the dynamics of student learning," Phys. Rev. ST–Phys. Educ. Res. **2**, 010103 (2006).
[6]R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66**(1), 64–74 (1998).
[7]R. Thornton and D. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," Am. J. Phys. **66**(4), 338–352 (1998).
[8]J. M. Bailey, "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, 2006.
[9]L. Ding, R. Chabay, B. Sherwood, and R. Beichner, "Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment," Phys. Rev. ST–Phys. Educ. Res. **2**, 010105-1–7 (2006).
[10]E. M. Bardar, E. E. Prather, K. Brecher, and T. F. Slater, "Development and validation of the light and spectroscopy concept inventory," Astron. Educ. Rev. **2**(5), 103–113 (2007).
[11]E. E. Prather, A. L. Rudolph, G. Brissenden, and W. M. Schlingman, "A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction," Am. J. Phys. **77**, 320–330 (2009).
[12]L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," Phys. Rev. ST–Phys. Educ. Res. **5**, 020103-1–17 (2009).
[13]W. M. Schlingman, E. E. Prather, C. S. Wallace, A. L. Rudolph, and G. Brissenden, "A classical test theory analysis of the light and spectroscopy

concept inventory national data set," Astron. Educ. Rev. **11**, 010107 (2012).
[14]D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).
[15]K. Williamson and S. Willoughby, "Student understanding of gravity in introductory college astronomy," Astron. Educ. Rev. **11**, 010105 (2012).
[16]A. Fraknoi, "Enrollments in astronomy 101 courses," Astron. Educ. Rev. **1**(1), 121–123 (2002).
[17]I. Halloun and D. Hestenes, "Common sense concepts about motion," Am. J. Phys. **53**(11), 1056–1065 (1985).
[18]I. Galili, "Interpretation of students' understanding of the concept of weightlessness," Res. Sci. Educ. **25**(1), 51–74 (1995).
[19]J. Dostal, "Student concepts of gravity," Masters thesis in Physics, Iowa State University, 2005.
[20]K. Williamson, S. Willoughby, and E. E. Prather, "Development of the Newtonian gravity concept inventory," Astron. Educ. Rev. **12**(1), 010107 (2013).
[21]J. Benson and F. Clark, "A guide for instrument development and validation," Am. J. Occup. Ther. **36**(12), 789–800 (1982).
[22]J. Creswell, *Qualitative Inquiry and Research Design*, 2nd ed. (Sage Publications, Thousand Oaks, CA, 2007).
[23]A. Strauss and C. Juliet, "Grounded theory methodology: An overview," edited by N. Denzin and Y. Lincoln, in *Handbook of Qualitative Research*, 1st ed. (Sage Publications, Thousand Oaks, CA, 1994), pp. 273–284.
[24]T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," Appl. Meas. Educ. **15**(3), 309–334 (2002).
[25]C. Wallace and J. M. Bailey, "Do concept inventories actually measure anything?," Astron. Educ. Rev. **9**, 010116 (2010).
[26]M. T. Kane, "An argument-based approach to validity," Psychol. Bull. **112**, 527–535 (1992).
[27]R. F. Gunstone and R. T. White, "Understanding of gravity," Sci. Educ. **65**, 291–299 (1981).
[28]M. Piburn, "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, 1998).
[29]R. E. Feeley, "Identifying student concepts of gravity," Masters thesis in Science and Teaching, The University of Maine, 2007.
[30]A. Asghar and J. C. Libarkin, "Gravity, magnetism and 'down': Non-physics college students' conceptions of gravity," Sci. Educ. **19**(1), 42–55 (2010).
[31]P. J. Mulvey and S. Nicholson, "Enrollments and degrees report," AIP Report, Number R-151-45 (2011).
[32]However, we do not claim that the classroom environments in this sample are representative of algebra-based college physics courses nationwide. Those instructors who voluntarily participated in this study are likely more involved in education reform efforts than average.
[33]Majors are self-reported, so any distinction between "Science, Engineering, or Architecture" and "Health or Medical Sciences" would be at the discretion of the students answering.
[34]F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores* (Addison-Wesley, Reading, MA, 1968).
[35]R. K. Hambleton and R. J. Jones, "Comparison of classical test theory and item response theory and their application to test development," Educ. Meas.: Issues Pract. **12**, 38–47 (1993).
[36]D. George and P. Mallery, *SPSS for Windows Step by Step: A Simple Guide and Reference* (Pearson Education, Boston, MA, 2009).
[37]This allows items that are *more* difficult to have a *higher* difficulty value. Note that this follows the work of Schlingman *et al*., which is opposite of the more conventional way of calculating item difficulty, where item difficulty is the fraction of *correct* responses.
[38]See supplementary material at http://dx.doi.org/10.1119/1.4945347 for histograms of the proportions of students choosing each answer choice for each of the 26 NGCI items, as well as interview notes from eleven "think-aloud" interviews with physics students.
[39]Interviews were with second-semester physics students just after they had finished first-semester physics. Audio recordings can be found in Ref. 42.
[40]R. N. Bolton and T. M. Bronkhorst, "Questionnaire pretesting: Computer-assisted coding of concurrent protocols," in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman (Jossey-Bass, San Francisco, 1996).
[41]R. Weathersby and R. Freyberg, *Study Guide and SPSS Manual: To Accompany Statistics for the Behavior Sciences* (Worth Publishers, New York, 2008).
[42]T. C. Urdan 2010, *Statistics in Plain English*, 3rd ed. (Taylor & Francis Group, New York, 2010).
[43]K. Williamson, "Development and calibration of a concept inventory to measure introductory college astronomy and physics students' understanding of Newtonian gravity," Doctoral thesis, Montana State University, 2013.

Standard Volumes

Sets of secondary volume measures were standard equipment for physics departments at the turn of the 20th century. This set, at Union College in Schenectady, New York, is probably made of pewter. (Notes and picture by Thomas B. Greenslade, Jr., Kenyon College)