The Development and Validation of the Planet Formation Concept Inventory (PFCI)

Molly N. Simon[a], Edward E. Prather[b], Sanlyn R. Buxner[c], and Chris D. Impey[b]

[a]University of Arizona Department of Planetary Sciences, 1629 E. University Blvd, Tucson, AZ 85721; [b]University of Arizona Department of Astronomy, Steward Observatory, Tucson, AZ; [c]University of Arizona, Department of Teaching, Learning, & Sociocultural Studies; Planetary Science Institute (PSI), Tucson, AZ

ARTICLE HISTORY
Compiled May 23, 2019

ABSTRACT
The discovery and characterization of thousands of planets orbiting distant stars has shed light on the origin of our own Solar System. It is important that college-level introductory astronomy students have a general understanding of the planet formation process before they are able to draw parallels between extrasolar systems and our own Solar System. In this work, we introduce the Planet Formation Concept Inventory (PFCI), an educational research tool used to assess student learning on the topic of planet formation. The PFCI Version 3 was administered to N = 561 students pre-instruction and N = 374 students post instruction. In this paper, we present a Classical Test Theory (CTT) analysis of the PFCI Version 3. Ultimately, we conclude that the PFCI is a reliable and valid instrument that can differentiate experts from novices, and can be used to assess college-level introductory astronomy students' learning on the topic of planet formation. Initial findings on class normalized gain scores indicate that the PFCI may be capable of assessing the effectiveness of different instructional models. In the future, we recommend a national study of the PFCI to discern its ability to provide insight regarding the ascribed and achieved characteristics of leaners and the effectiveness of different instructional strategies being used to teach this topic. The results from a national study could ultimately lead to the development of more targeted active learning strategies capable of helping learners significantly increase their conceptual knowledge and reasoning abilities on this topic.

KEYWORDS
assessment, undergraduate, astronomy, planet formation, reasoning difficulties, concept inventory, instrument development

1. Introduction

The topic of planet formation has become increasingly relevant to the introductory astronomy curriculum as the characterization and discovery of thousands of planets outside of our Solar System (exoplanets) has become one of the most active areas of new research for the discipline. Teaching students about the formation of our Solar System (and the process of planet formation on a more general scale) lends to a better understanding of the origin and evolution of exoplanetary systems more generally.

CONTACT Molly N. Simon Email: msimon@lpl.arizona.edu

As a result, it is imperative that introductory college students have a preliminary understanding of planet formation before they are able to draw comparisons between the thousands of newly discovered solar systems, and our own Solar System.

A concept inventory is a multiple-choice style instrument that addresses a single topic or closely related set of topics. Concept inventories items should be written in a way that emulates students' natural language, and, as a result, scientific jargon should be minimal. Most importantly, concept inventories can be distinguished from traditional multiple-choice tests in that they use students' pre-instructional ideas as the basis of their distractors (incorrect answer choices) (**?**). Concept inventories are particularly useful for assessing students' pre and post-instructional conceptual understanding of a specific topic, and the efficacy of pedagogy developed to teach that topic (e.g. **?**; **?**). In astronomy and planetary science, concept inventories have already been developed on the topics of: stars and their properties (**?**), the greenhouse effect (**?**), light and spectroscopy (**?**), lunar phases (**?**), and Newtonian gravity (**?**).

The Planet Formation Concept Inventory (PFCI) explores the many conceptual and reasoning difficulties students experience when learning about planetary origins (**?**; **?**). The PFCI was developed, in particular, to assess students' understanding of planet formation in general education undergraduate astronomy and planetary science courses, typically referred to as "ASTRO 101." Prior research has shown that students come into their ASTRO 101 courses with a variety of pre-instructional ideas on the evolution of planetary systems. These ideas can interfere with their ability to develop a comprehensive understanding of the topic. An in depth explanation of these pre-instructional ideas related to planet formation can be found in **?** and the references therein.

The development and validation of the PFCI was motivated not only by the topic of planet formation's aforementioned relevance to ASTRO 101 curriculum, but also due to the fact that the astronomy education research literature is incomplete when it comes to studies that reflect our understanding of students' perceptions on planetary origins (e.g. **?**). The PFCI is the first concept inventory that addresses topics central to an understanding of planet formation. Through the work reported here, we will quantitatively demonstrate that the PFCI is a reliable and valid instrument that can assess students' change in understanding on the topic of planet formation as a result of instruction in ASTRO 101.

2. Methods

2.1. Concept Domain

The PFCI's range of topics (concept domain) was determined after an analysis of course materials from 34 undergraduate courses that covered planet formation as part of their curriculum; as well as after a thorough analysis of students' pre-instructional ideas (**?**). The concept domain covered by the PFCI addresses the following topics:

- the physical composition of the planets in our Solar System
- condensation temperature - and the role the condensation of elements plays in determining the physical characteristics of the planets in our Solar System
- the accretion process during planet formation
- planetary orbits and migration
- definitions of a planet, exoplanet, and solar system
- the nebular theory (and the fact that the formation of the Universe and the forma-

tion of our Solar System are independent events)

Each of these topics are represented by at least two questions on the PFCI.

2.2. Test Question/Item Development

To evaluate student learning of planet formation most effectively and for ease of scoring and analysis, we selected a multiple-choice format for the PFCI. Again, all of the items on the PFCI originated from prior research into students' understanding of planet formation before instruction (**?**; **?**). When developing each of the multiple-choice items on the PFCI, we commonly referred back to the 31 item writing guidelines for classroom assessment described in **?**. As a result, each of the multiple-choice items consisted of a question stem followed by either four or five answer choices. Each item had a clearly worded correct answer, and the rest of the answer choices consisted of common student naive ideas on the topic being addressed, and served as distractors. We wrote each of the item distractors in students' natural language, and limited scientific jargon as much as possible. After the original set of test items were developed, they were reviewed by faculty members whose research is planet formation-focused and who have experience teaching planet formation in general education undergraduate introductory courses.

2.3. Preliminary Versions of the PFCI

The PFCI was developed using an iterative design process over 4 semesters. The first version of the PFCI consisted of 20 multiple-choice questions and 3 demographic questions (major, gender, previous instruction on planet formation). It was administered to six introductory astronomy and planetary science courses (N = 455 students) at The University of Arizona at the beginning of the Fall 2017 semester before any relevant material was taught (this version was only administered as a pre-test). Students enrolled in these courses were overwhelmingly non-majors fulfilling their college's natural science requirement.

For the first administration of the PFCI, we broke the 20 content questions into three "mini" concept inventories. Students (at random) either answered questions 1-7, 8-14, or 15-20 (along with the three demographic questions). Students were instructed not to put their names on the concept inventory in order to preserve their anonymity. In addition to selecting an answer choice, students were asked to provide 1-2 sentences explaining why they selected a particular answer. Students were also encouraged to provide feedback regarding the clarity of the question stem and the different answer choices. This multiple-choice with explanation of reasoning (MCER) approach allowed us to determine whether the students were interpreting the questions on the PFCI as we intended. Our analysis of students' MCER responses also allowed us to better determine the clarity and quality of the instrument items, and whether these items were well matched to students' preinstructional ideas. We did not perform a statistical analysis of the PFCI Version 1 since it was broken into three mini concept inventories. Instead, we coded students' MCER responses using the same post-hoc coding method described in detail in **?**. The MCER responses did not lead to any additional content codes from those uncovered in **?**, but our analysis of student responses did lead to the revision of seven items and the creation of one additional item. These revisions and the addition of a 21^{st} item constituted what became the second version of the PFCI.

Version 2 was administered to two ASTRO 101 courses (N = 141 students) during the Spring 2018 semester. This version was administered in its entirety in typical multiple-choice format at the end of the semester (just post-test). Once again, students were instructed not to put their names on the concept inventory in order to preserve their anonymity. The second version of the PFCI consisted of 21 content items, and the same 3 demographic questions from Version 1. The average score on the PFCI Version 2 was 10.4/21 (49%) with a standard deviation of 3.6 (17%). We performed a brief statistical analysis on this version, including calculating the instrument Cronbach's alpha (α) = 0.681. Item difficulty values (p) for the PFCI Version 2 ranged from 0.18 - 0.81 with an average $p = 0.49$. Item discrimination (ρ_{pbis}) values ranged from 0.01 - 0.43 with an average $\rho_{pbis} = 0.25$. A detailed explanation of these statistical tests and their interpretations can be found in Sections **??** and **??**. The results of our statistical analysis were used to inform the revision of seven test items. Five items were revised due to low item discrimination values, one item was revised due to a low item difficulty value (indicating the question was too difficult even after adequate instruction), and an additional item was revised due to poor item discrimination *and* item difficulty values.

At this phase of our iterative revision process, we solicited the feedback of three planetary science professors, one astronomy professor, and three science education researchers. The planetary science/astronomy professors provided feedback on the questions' scientific accuracy. The education researchers analyzed the items in the instrument, and provided feedback to ensure that the topics covered and language used were appropriate for undergraduate ASTRO 101 students. The combination of our statistical analysis and in-depth feedback from the science professors and education researcher specialists ultimately lead to the revision of 15 of the PFCI's 21 items. A majority of the revisions were relatively minor with the exception of one item, which was removed entirely. The completion of these revisions lead to Version 3 of the PFCI.

Version 3 of the PFCI consisted of 20 content items and 3 demographic items. It was *this* version that was administered to ASTRO 101 students pre and post-instruction and underwent an in-depth reliability, validity, and item analysis as we will describe in Sections **??** - **??**. The final and most robust version of the PFCI Version 3 can be found in Appendix **??**.

3. Results

We performed an in-depth quantitative analysis of the PFCI Version 3 using methods consistent with classical test theory (CTT). CTT is typically utilized as a means to determine the statistical properties of test items and, if necessary, eliminate or revise the test items that do not meet pre-established criteria. CTT methods also outline a procedure to conduct a statistical analysis of an instrument's reliability and validity (**?**; **?**). The results from our CTT analysis of the PFCI's Version 3 are presented in Sections **??** - **??**.

3.1. PFCI Version 3 Sample

Version 3 of the PFCI was administered during the Fall 2018 semester to seven ASTRO 101 courses before any relevant material was taught (pre-test), and to six of those same classes *again* at the end of the semester (post-test). Each student received the same version of the PFCI, and students recorded their answers directly onto the instruments. Students' scores were recorded manually and double checked by education researchers. For our analysis of the PFCI Version 3, we removed any students from the dataset who selected the same answer for all 20 questions or who answered the questions in a specific visual pattern - indicative of the fact that they did not earnestly answer the items of the PFCI. Following the procedure outlined in **?**, we also removed data from students who left more than two questions blank in order to avoid early question bias. We then matched students who took the PFCI Version 3 both pre and post (matched pairs). The number of test takers can be found in Table **??**.

3.2. Item Analysis: Item Difficulty & Discrimination

Item difficulty, p, is defined as the proportion of students who answered a specific question correctly (**?**). As a result, items with lower difficulty values are considered *more difficult* than items with higher difficulty values. According to **?**, the range of acceptable p-values is typically between $0.2 < p < 0.8$. Our pre-test item difficulty values ranged from 0.17-0.81 with an average $p = 0.46$. After instruction (post-test), our item difficulty values ranged from 0.34-0.81, with an average $p = 0.58$ (see Table **??**). Ideally, students should preform better on PFCI items after targeted instruction on these topics. As a result, item difficulty values should be higher (a greater proportion of students answered the item correctly) for all of the PFCI's post-test items. This was true for all but two items (#10 and #12) as shown in Figure **??**. We will discuss these items, along with items #1, #3, and #13 (who pre or post-test item difficulty values were slightly outside of the desirable range) in Section **??**.

Item discrimination is used to measure how effectively an item differentiates between test takers who do well and those who do poorly on the entire test. It is defined using an item's point biserial,

$$\rho_{pbis} = \frac{(\mu_{+} - \mu_{X})}{\sigma_{X}} \cdot \sqrt{\frac{p}{q}} \tag{1}$$

where μ_{+} is the mean test score for those who answered the question correctly, μ_X is the mean test score for the entire sample, σ_X is the standard deviation of all of the test scores, p is the item difficulty, and q = (1-p). Values of ρ_{pbis} can range from -1.00 to +1.00 with a value of 0 indicative of zero correlation. A positive ρ_{pbis} value indicates that there is a positive correlation between the item score and the test score overall, thus, students scoring higher on the exam were more likely to answer that particular item correctly when compared to students whose test scores were low. This indicates that the item successfully discriminated between high- and low-scoring students (**?**). Similarly, a negative ρ_{pbis} value indicates that students who are performing well on the exam are answering that item *incorrectly* more often than students performing poorly on the exam. Items with very low or negative discrimination values are poor discriminators of student understanding.

For the post-test administration of the PFCI Version 3, our ρ_{pbis} values ranged from 0.04-0.45 with an average of 0.31 (see Table **??**). Our average ρ_{pbis} value was consistent with other concept inventories within this topical domain and with this population of test-takers (see e.g. **?**; **?**; **?**; **?**). The minimum desired point biserial value is 0.2, with an ideal range between 0.3-0.7 (**?**; **?**). All but three of our items had point biserial values within the ideal range. Items with point biserial values less than 0.20 can still be considered acceptable if the point biserial coefficient is two standard deviations above 0.00 and the sample size (N) is ≥ 50. The standard deviation is defined as:

$$\sigma_\rho = \frac{1}{\sqrt{N-1}} \tag{2}$$

(**?**). For a sample size of N = 374, the standard deviation of the point biserial coefficient is 0.052, and two standard deviations is 0.104. All of our items except #4 and #9 were above ρ_{pbis} = 0.104 (see Figure **??**). These two items underwent additional review to determine whether they should be discarded or modified before publishing the final version of the PFCI. A detailed explanation of our review process can be found in Section **??**.

3.3. Student Learning Gains

To measure average learning gains, we first calculated normalized gain scores for each individual student ($g_{student}$) in the matched-pairs dataset (N = 287 students) (**?**):

$$g_{student} = \frac{(post\%) - (pre\%)}{100 - (pre\%)} \tag{3}$$

A paired-sample t-test comparing individual students' pre and post-test scores showed that students' scores after instruction were significantly higher than their scores on the pre-test, t(286) = -14.426, p <0.001. Figure **??** shows that 76% of students performed better on the post-test, indicating that the overwhelming majority of students demonstrated a better understanding of planet formation after instruction.

As a next step, we used the student data to calculate average normalized gain score for each class, g_{class}. **?** and **?** defined three categories of gain: low normalized gain (g < 0.3), medium normalized gain (0.3 < g < 0.7), and high normalized gain (g > 0.7). We followed the same standard when analyzing the normalized gain scores for each of the six classes we surveyed. The mean (M), standard deviation (SD), and class normalized gain scores are reported in Table **??**.

Similar to **?**, we found that the range of normalized gain scores achieved by individual students (-0.875 < $g_{student}$ < 1.0) was much greater than the range of gain scores for the six classes (0.170 < g_{class} < 0.377). Over one-third of students (38%) scored within the medium gain range (0.3 - 0.7), and an additional 4% of students had normalized learning gain scores in the high range ($g_{student}$ > 0.7) as shown in Figure **??**. We explored the possibility that higher gain scores could be attributed to college major, gender, previous planet formation exposure, or type of instruction in Section **??**.

3.4. Instrument Reliability

We assessed the reliability of the PFCI by calculating Cronbach's alpha (α). Cronbach's alpha is defined using the following formula:

$$\alpha = \frac{K}{K-1} \cdot (1 - \frac{\sum \sigma_i^2}{\sigma_x^2}) \tag{4}$$

where K = the number of test items, σ_i^2 is the variance of each item, and σ_x^2 is the variance of the entire test (?). A reliability coefficient of 0.70 or higher is considered acceptable (?). We used the sample of matched pairs (N = 287 students) to calculate the PFCI's internal consistency both before and after instruction. Before instruction, $\alpha = 0.658$, which was slightly lower than what is deemed acceptable. For the post-test, however, $\alpha = 0.726$. This increase in α indicates that students answer more reliably to the items on the PFCI after instruction. Furthermore, this increase in α post instruction demonstrates that Version 3 of the PFCI is a reliable instrument for assessing student understanding of this topic area.

3.5. Instrument Validity

Validity refers to how well a scientific test (and the test items within) measure what the instrument intends (?). Following the procedure outlined in ?, we evaluated three types of validity for the PFCI: content validity, face validity, and concurrent validity.

3.5.1. Content Validity

Content validity measures the extent to which the test questions are scientifically accurate - and whether or not the test items are representative of the concept domain the instrument seeks to evaluate (?). To evaluate the PFCI's content validity, we solicited the help of three planetary science professors, one astronomy professor, and three science education researchers (see Section ??). Their suggested language/content changes were implemented before Version 3 was administered. Overall, the professors assessing the PFCI were in strong agreement that the instrument was well written and scientifically accurate, indicative of satisfactory content validity.

3.5.2. Face Validity

Face validity is utilized to determine whether the concept domain addressed by the instrument covers the topics most appropriate for assessing students' knowledge (?). To assess the face validity of the PFCI, we conducted an analysis of course syllabi (and lecture slides, when available) from 34 undergraduate introductory astronomy and planetary science courses that covered the topic of planet formation (refer to ?). Our investigation supported the claim that the PFCI does have face validity, and our instrument addresses the planet formation content most commonly taught in ASTRO 101 courses.

3.5.3. Concurrent Validity

Concurrent validity is an instrument's ability to distinguish between distinct populations of test takers (?; ?). To evaluate the PFCI's concurrent validity, we administered the PFCI to a set of 12 planetary science graduate students, and two postdoctoral researchers (M = 19.14/20, 95.7%, SD = 0.949/20, 4.74%). We performed a t-test to determine whether the PFCI was able to measure a statistically significant difference in post-test scores between the graduate students/post-docs (experts) and the undergraduates (novices). An independent sample t-test confirmed that the PFCI was successfully able to distinguish between the two test taking populations, $t(386)$

= -7.215, p < 0.001.

Due to the fact that Version 3 of the PFCI satisfies the criteria outlined for content, face, and concurrent validity, we conclude that our instrument can be used in a classroom setting to distinguish between different levels of understanding amongst our target population on the topic of planet formation.

4. Discussion

4.1. Items Requiring Justification

As mentioned in Section **??**, items #1 and #13 had lower than desirable p-values on the pre-test. Their corresponding post-test p-values were within the acceptable range, and therefore these items were retained in the final version of the PFCI. Additionally, items #3 and #12 had p-values on the cusp of indicating that these questions were 'too easy.' This was unsurprising, however, since these items were developed to act as the instrument's 'baseline.' Since these items had acceptable point-biserial values, they were unchanged in the PFCI's final version.

Ideally, item difficulty values should be higher (a greater proportion of students answered the item correctly) for all of the PFCI's items after instruction. We observed this trend for all but two items (#10 and #12) as shown in Figure **??**. Since item #10 had such a minimal decrease in p-value, and since its point biserial value was within the acceptable range, we retained item #10 in the final version of the PFCI Version 3. As mentioned previously, we did not modify item #12 because it covered the most basic definition of a planet, the baseline of student knowledge for the PFCI.

Items #4 and #9 were flagged for having ρ_{pbis} values outside of the desirable range. These items were originally included in the PFCI because they addressed known student learning difficulties on the topics of planetary migration and condensation temperature, respectively.

On Version 3 of the PFCI, item #4 read as follows:

4. Which describes how the locations (relative to the Sun) of the planets in our Solar System may have changed over time?

A. They changed because space is constantly moving and expanding

B. They changed because the planets are constantly colliding with each other

C. The larger planets may have changed locations early in the Solar System's history because of the gravitational interactions between them

D. The locations of the planets have not changed over time; they formed in the same locations they are in now

The correct answer choice is C. Before instruction, 40.1% of students picked answer choice A, and 39.9% selected answer choice C. After instruction, however, only 24.1% of students picked answer choice A, and 55.1% chose answer choice C. The results of an independent sample t-test showed that the percentage of students who answered #4 correctly after instruction was significantly higher than the corresponding percentage before instruction, $t(933)$ = -4.600, p < 0.006.

Since we saw a statistically significant increase in the percentage of students selecting the correct answer after instruction, we assert that this item measures a concept (giant planet migration) that targeted instruction on this topic is able to address. The fact that the ρ_{pbis} value is low simply indicates that while some students *do* learn the concept covered in this particular item, many of the students who answer this item correctly are still struggling with content related to other items on the instrument. A low ρ_{pbis} value could also be indicative of a test item that is being answered incorrectly by high performing students. Due to the increase in the percentage of students answering this item correctly after instruction, we did not modify or remove item #4 from the final version of the PFCI.

Item #9 was flagged due to its less than acceptable ρ_{pbis} value compounded with its nearly negligible increase in item difficulty post-instruction.

Item #9 read as follows:

9. In the outer Solar System, which materials were able to become a solid?
   A. Only metals (e.g. iron)
   B. Only rocky (silicon-based) minerals
   C. Only hydrogen compounds (e.g. water and ammonia)
   D. Metals, rocky minerals, and hydrogen compounds
   E. Neither metals, silicon based minerals, nor hydrogen compounds

The correct answer choice is D. Prior to instruction, 16.6% of students selected C while 64% chose D. After instruction, the numbers remained relatively unchanged, with 17.9% of students picking C and 65.5% selecting D. The intent of this item was to measure whether or not students had a robust understanding of the role the condensation of elements plays in determining the compositional differences between the terrestrial and jovian planets. As a result, we developed a question with distractors more aligned with known student reasoning difficulties on the topic of condensation temperature/planetary composition as described in **?**).

We propose the following question as a replacement for item #9 in the final version of the PFCI Version 3:

9. Jupiter, Saturn, Uranus, and Neptune (the outer planets) were able to grow much larger than Mercury, Venus, Earth, and Mars because:
   A. In the locations where the outer planets formed metals, rocky minerals, and icy minerals were all able to solidify. As a result, all of these materials could be used to form the outer planets.
   B. The gravitational force far from the Sun was much weaker, allowing the outer planets to grow to much larger sizes.
   C. In the outer Solar System there was much more rocky material than icy material. This made it possible for the outer planets to attract their large gaseous envelopes.
   D. During the Solar System's formation, the Sun ejected additional solids into the outer Solar System. These solids were eventually used to form the outer planets.

The new correct answer choice is A.

4.2. Further Exploration of Learning Gains

Section **??** illustrated that the range of normalized gain scores achieved by individual students ($g_{student}$) was very wide. As a result, we explored the possibility that higher gain scores could be attributed to college major, previous instruction on planet formation, gender, or type of course instruction (e.g. interactive versus lecture based). Of the 109 students (38% of the entire matched-pairs sample) whose normalized gain scores were within the 'medium gain' range, 15.6% were science majors, the gender distribution was approximately 50-50, and 32.1% of these students had learned about planet formation previously in some capacity (either in high school or college). For the small sample of students (N = 11) whose normalized gain scores were within the 'high' range, 54.5% were science majors, 63.7% were female, and 45.4% had previously learned about planet formation. Histograms of these demographic breakdowns for the entire matched pairs sample as well as the medium and high gain subsamples can be found in Figure **??**.

Due to the increase in the percentage of science majors between the medium and high normalized gain subsamples, we decided to explore whether or not the PFCI could differentiate between science and non-science majors for the entire matched pairs dataset. An independent sample t-test confirmed that science majors' average normalized gain scores ($< g > = 0.38$) were significantly higher than those of non-science majors ($< g > = 0.21$), t(275) = 3.973, p <0.001. The same result could not be found, however, between male and female test-takers or between students who had previous instruction on planet formation and those who did not. The normalized gain scores between those demographic populations were statistically indistinguishable. It was particularly perplexing that the instrument was not biased towards students with previous exposure to planet formation. This finding could be indicative of the fact that planet formation is not adequately taught at the high school or introductory college level, and that there may be a discrepancy between the concepts covered in students' previous courses and those emphasized on the PFCI.

We also explored the possibility that type of course instruction (interactive versus lecture based) played a role in helping students achieve higher normalized gain scores (**?**). Typically, implementing active learning strategies into the ASTRO 101 classroom (such as Lecture Tutorials or Think-Pair-Share/peer instruction questions), allows students to "explore the reasoning behind their answers. In doing so, they improve their reasoning skills and their understanding of core topics. Systematic studies have shown that [active learning] strategies can improve students' understanding by two full letter grades beyond what traditional lectures accomplish" (**?**, p. 44).

In order to evaluate how interactive the six courses were, we administered a questionnaire at the end of the semester that asked the course instructors to describe what active learning strategies (if any) they used to supplement traditional lecture when teaching planet formation. Of the six courses surveyed, three were classified as 'interactive,' meaning the instructors used two interactive learning strategies when teaching planet formation, two were classified as 'partially interactive' meaning the instructors used one active learning strategy, and one course was classified as 'lecture only', since the instructor indicated that no active learning strategies were used when teaching planet formation in their course. The interactive learning strategies included the use of Lecture-Tutorials and peer instruction/Think-Pair-Share. Table **??** shows the course number, level of interactivity, and class normalized gain scores for each of the six courses.

It was unsurprising that two of the three classes with the highest average gain scores were classified as 'Interactive' and the course with the lowest average gain score utilized only traditional lecture. Unfortunately, however, an independent sample t-test used to compare students' individual gain scores could not distinguish between students in interactive classes, and those in partially interactive or lecture only classes. Although implementation of active learning strategies can lead to higher class averaged normalized gain scores, research has shown that a faculty member's successful implementation of these strategies can be the most critical factor in determining class gain score (**?**). Instructors may not implement these strategies successfully into their specific classroom context. Furthermore, if an active learning strategy *is* implemented successfully, it may address only a small portion of what is covered by the PFCI's content domain. Instructors were not given a copy of the PFCI until after the post-test was administered, as we did not want them to teach towards the test. Since planet formation covers a broad array of concepts, there is no uniformly accepted way to teach it, and according to the instructor questionnaires, basic definitions (planet, exoplanet, dwarf planet), planetary motion/migration, and the origin of the Universe (Big Bang) were not covered by 50% of the courses we surveyed. Based on these findings, we recommend re-analyzing the PFCI's ability to distinguish between interactive and lecture-only courses with a significantly larger sample size of participating courses, and with courses that explictly teach topics with a greater amount of content overlap with the PFCI.

5. Conclusions & Future Work

In this work, we have provided the science education community with a detailed description of the Planet Formation Concept Inventory's (PFCI) development process. We have also utilized Classical Test Theory (CTT) to conduct a thorough statistical analysis of the PFCI's reliability, validity, and item statistics. Based on our findings, we conclude that Version 3 of the PFCI is a reliable and valid instrument that can measure change in ASTRO 101 students' conceptual understanding of planet formation before and after instruction.

There are several research avenues to explore based upon the results of this project. First and foremost, it will be beneficial to conduct a nationwide study using the PFCI. During such a national study, it will be valuable to investigate the instrument's reliability and validity when administered to students at different types of academic institutions with a larger sample of course instructors and instructional models. With a more generalizable dataset, we could revisit whether or not the PFCI is able to differentiate between courses where active engagement is prevalent and those where instructors use predominantly lecture-based practices (e.g. **?**).

Another avenue for future research would be to corroborate our CTT findings through an analysis of the PFCI using Item Response Theory (IRT). Item response theory has the ability to provide a more in-depth analysis of each of the PFCI's content items, and we plan to utilize *both* IRT and CTT when analyzing the data from the aforementioned national study (see e.g. **?**).

The results of a national study could also potentially shed light on specific cases where instruction is not resulting in greater student understanding. By looking at a more robust sample of post-test results, we would be able to identify and address specific conceptual and reasoning difficulties that are resistant to change even after targeted instruction. These findings would allow us to develop new active learning strategies that explicitly target persistent student conceptual and reasoning difficulties.

The PFCI will assess the efficacy of these active learning strategies, which can be utilized to promote a much deeper degree of learning on the topic of planet formation and the subtopics therein.

Disclosure statement

This is a confirmation that for all authors no competing financial or other interests exist.

Funding

This work was supported through a NASA traineeship grant awarded to the Arizona/NASA Space Grant Consortium.

6. References

7. Appendices

Appendix A. Final Version of the PFCI Version 3

The final version of the PFCI is presented here. This version is the same as Version 3 with the exception of item #9, which was revised so as to include distractors more aligned with known student reasoning difficulties (see Section **??**).

Final Version of the PFCI

Directions:

Read the entire question and ALL of the answer choices carefully before selecting an answer. Fill in the bubbles completely! Each question has *one* correct answer unless otherwise specified.

Instead of writing your name on this form, fill in the numerical bubbles with the ***last four digits of your phone number*** (we will not be able to look you up with this information, and your identity will remain anonymous).

Enter last four digits below:

⓪	⓪	⓪	⓪
①	①	①	①
②	②	②	②
③	③	③	③
④	④	④	④
⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨

1. How did the planets in our Solar System form?

O They formed from a collision between our Sun and a nearby star

O They formed from the energy and matter released at the same time as the Big Bang

O They formed from the collapse of a cloud composed of gas and dust

O They formed from the remains of a massive stellar (star) explosion

O They formed from material that was pulled in from a nearby solar system by our Sun

2. Which of the following statements is *FALSE?*

O A planet must orbit around a star (sun)

O A planet must have an atmosphere

O A planet must clear its orbit of surrounding debris

O A planet must be roughly spherical in shape

3. In our Solar System, what best describes the physical characteristics of Mercury, Venus, Earth, and Mars?

O These planets are dense, small, and are made primarily of hydrogen and helium gas

O These planets are dense, small, and are made primarily of rocks and metals

O These planets are dense, large, and are made primarily of rocks and metals

O These planets have low densities, are large, and are made primarily of hydrogen and helium gas

O These planets have low densities, are large, and are made primarily of icy material

4. Which describes how the locations (relative to the Sun) of the planets in our Solar System may have changed over time?

O They changed because space is constantly moving and expanding

O They changed because the planets are constantly colliding with each other

O The larger planets may have changed locations early in the Solar System's history because of the gravitational interactions between them

O The locations of the planets have not changed over time; they formed in the same locations they are in now

5. Which of these objects would you expect to find in our Solar System? (*CIRCLE ALL THAT APPLY*)

O Comets

O Asteroids

O Dwarf Planets

O Exoplanets

O The Milky Way Galaxy

6. Which of these scenarios best describes the general planet formation (accretion) process?

O Dust grains continuously collide, accumulate more mass, and develop into planets

O Material is ejected into the Solar System from an explosion, and this material forms the planets

O Once the planets grow large enough, the Sun's gravity causes the growing planets to accumulate all of the matter around them

O A pressure build up in the solar nebula due to the birth of our Sun leads to the formation of the planets

O At the time of the Big Bang, material collides, grows in size, and forms planets

7. For our Solar System, which statement is *TRUE* regarding which planet[s] completed their formation process first?

O Mercury formed first because it is the closest to the Sun, and Neptune formed last because it is the farthest away

O The rocky planets all formed together first, and then the gas giant planets started their formation millions of years later

O The gas giant planets all formed together first, and then the rocky planets started their formation millions of years later

O Neptune formed first because it is the farthest from the Sun, Mercury formed last because it is the closest to the Sun

O All of the planets in our Solar System formed at approximately the same time

8. What is the definition of an exoplanet?

O A planet at the edge of the Solar System

O A planet outside of our Solar System

O A planet no longer bound by gravity to its star

O A planet that is habitable

O A planet that does not clear its orbit of surrounding debris

9. Jupiter, Saturn, Uranus, and Neptune (the outer planets) were able to grow much larger than Mercury, Venus, Earth, and Mars because:

O In the locations where the giant planets formed metals, rocky minerals, and icy minerals were all able to solidify. As a result, all of these materials could be used to form the outer planets

O The gravitational force far from the Sun was much weaker, allowing the outer planets to grow to much larger sizes

O In the outer Solar System there was much more rocky material than icy material. This made it possible for the outer planets to attract their large gaseous envelopes

O During the Solar System's formation, the Sun ejected additional solids into the outer Solar System. These solids were eventually used to form the outer planets

10. During the planet formation process, what is the primary role of the force of gravity?

O Gravity helps bodies with enough mass attract surrounding dust and gas so they can continue to grow into planets

O Gravity is the force that causes denser, more massive planets to form closer to the Sun

O Gravity determines which material (e.g. metals and gas) will be prevalent at certain distances from the Sun

O Gravity keeps the growing planets from collapsing on themselves if they get too massive during the formation process

11. In our Solar System, what best describes the physical characteristics of Jupiter, Saturn, Uranus, and Neptune?

O These planets are dense, small, and are made primarily of hydrogen and helium gas

O These planets are dense, small, and are made primarily of rocks and metals

O These planets are dense, large, and are made primarily of rocks and metals

O These planets have low densities, are large, and are made primarily of gases and icy material

O These planets have low densities, are large, and are made of strictly hydrogen and helium gas

12. Which of these most accurately describes a planet?

O A planet orbits around another larger, rocky body

O A planet orbits around a star

O A planet is an object that is massive enough to fuse hydrogen into helium

O A planet must be able to sustain life

13. Why do the planets in our Solar System orbit the Sun in the same plane?

O The planets were ejected into this configuration at the time of the Solar System's formation

O The planets formed from a flattened disk-like structure, which caused the planets to orbit in this configuration

O The planets orbit around the Sun on retrograde orbits, and these orbits require the planets to be in the same plane

O The planets were pulled into this configuration by the gravity of nearby asteroids and comets

14. In our Solar System, why did rocky planets form close to the Sun while the gaseous planets formed further away?

○ Close to the Sun, gravity was only strong enough to pull the rockier planets close in

○ Close to the Sun, planets composed of mainly gas were incapable of remaining stable

○ Close to the Sun, only heavy elements (like rocks and metals) could solidify at such high temperatures and eventually form a planet

○ Close to the Sun, all of the gaseous material was used to create the young Sun, so there was no material left to form the gas planets close in

15. The planets in our Solar System orbit the Sun in ____ direction[s], at ____ speed[s], and on ____ orbits.

○ The same, different, elliptical

○ The same, the same, elliptical

○ Different, the same, circular

○ Different, different, elliptical

○ The same, different, circular

16. Which of these best describes how the composition of the planets in our Solar System changes with *increasing* distance from the Sun?

○ Rocky Planets → Gas Planets → Icy Planets

○ Gas Planets → Rocky Planets → Icy Planets

○ Icy Planets → Rocky Planets → Gas Planets

○ Rocky Planets → Icy Planets → Gas Planets

○ Gas Planets → Icy Planets → Rocky Planets

17. When did our Solar System form relative to the Big Bang?

O Our Solar System formed before the Big Bang

O Our Solar System formed at the same time as the Big Bang

O Our Solar System formed immediately after the Big Bang

O Our Solar System formed a long time after the Big Bang

18. The discovery of exoplanetary systems has supported the idea that:

O Exoplanetary systems look exactly like our Solar System, with a combination of rocky and gaseous planets

O Exoplanetary systems look entirely different than our Solar System, with planets made of materials not found in our Solar System

O Exoplanetary systems must have Jupiter-sized planets orbiting close to their host stars

O Exoplanetary systems are different from our Solar System in that every planet discovered in these systems has the potential for life

O Exoplanetary systems are likely similar to our Solar System in terms of the general formation process, but the locations and compositions of the planets may be different

19. What is the definition of a dwarf planet?

O A planet at the edge of a solar system

O A planet outside of our Solar System

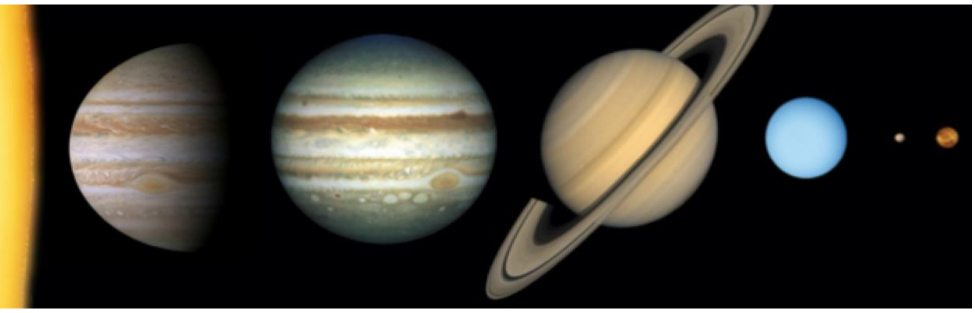
O A planet with an irregular orbit around a star

O A planet that is habitable

O A planet that does not clear its orbit of surrounding debris

20. Below is a depiction of a hypothetical solar system. Based on the image below, which of the following answers correctly describes why the planets are in the locations shown?

NOTE: The sizes of the planets are not to scale relative to the distances between them.



O All of the planets in this solar system formed exactly where they are shown

O The largest planets in this solar system moved inward during the formation process due to planetary migration

O The largest planets in this solar system moved inward during the formation process because the Universe is constantly moving and expanding

O The strong gravitational pull of the star caused the large and small planets to switch positions

21. Which of the following best characterizes your academic major(s)?

O Science major (e.g. physics, chemistry, biology)

O Non-science major (e.g. history, business, dance, etc…)

O Double major

O Undecided

O Other

22. What gender do you identify with?

O Male

O Female

O Non-binary

O Non-conforming

O Other

23. Have you ever taken a course besides this course that covered the topic of planet formation?

O Yes, in high school

O Yes, at a 4-year college

O Yes, at a community college

O Yes, other

O No

Table 1. PFCI Version 3 Participant Distribution

Administration	Number of Students (N)
Pre-Test	561
Post-Test	374
Matched-Pairs	287

Table 2. PFCI Version 3 Item Statistics

Item Number	**Pre-*p***	**Post-*p***	ρ_{pbis}
	(N=561)	(N=374)	(N=374)
1	**0.17**	0.52	0.35
2	0.44	0.64	0.33
3	0.64	**0.81**	0.45
4	0.40	0.55	**0.08**
5	0.37	0.40	0.37
6	0.37	0.51	0.44
7	0.26	0.34	0.22
8	0.49	0.56	0.20
9	0.64	0.66	**0.04**
10	0.49	0.47	0.34
11	0.65	0.72	0.39
12	**0.81**	0.80	0.41
13	**0.19**	0.60	0.37
14	0.49	0.64	0.23
15	0.53	0.68	0.26
16	0.67	0.71	0.42
17	0.48	0.49	0.45
18	0.50	0.59	0.36
19	0.30	0.44	0.32
20	0.27	0.41	**0.14**

Item difficulty (p) pre and post-test values and item discrimination (ρ_{pbis}) post-test values for the PFCI Version 3. Values outside of the conventionally accepted range are in bold.

Table 3. Measured Learning Gains

Course Number	**# of Students**	**Pre-Test** $< M >$ %	**Pre-Test** $< SD >$ %	**Post-Test** $< M >$ %	**Post-Test** $< SD >$ %	g_{class}
PTYS 206	24	48.1	19.6	61.5	20.2	0.295
PTYS 214	9	59.4	27.2	77.2	16.2	0.377
PTYS 170B2 (1)	52	43.8	17.1	58.1	20.5	0.269
PTYS 170B2 (2)	40	51.0	14.9	60.1	20.2	0.170
PTYS 170A1	92	48.0	16.0	60.4	15.1	0.227
ASTR 170B1	70	41.5	15.7	53.0	20.5	0.203
Whole Sample	**287**	**46.4**	**17.1**	**58.7**	**19.1**	**0.231**

Table 4. Course Interactivity

Course Number	Interactivity Level	g_{class}
PTYS 214	Interactive	0.377
PTYS 206	Interactive	0.295
PTYS 170B2 (1)	Partially Interactive	0.269
PTYS 170A1	Partially Interactive	0.227
ASTR 170B1	Interactive	0.203
PTYS 170B2 (2)	Lecture Only	0.170

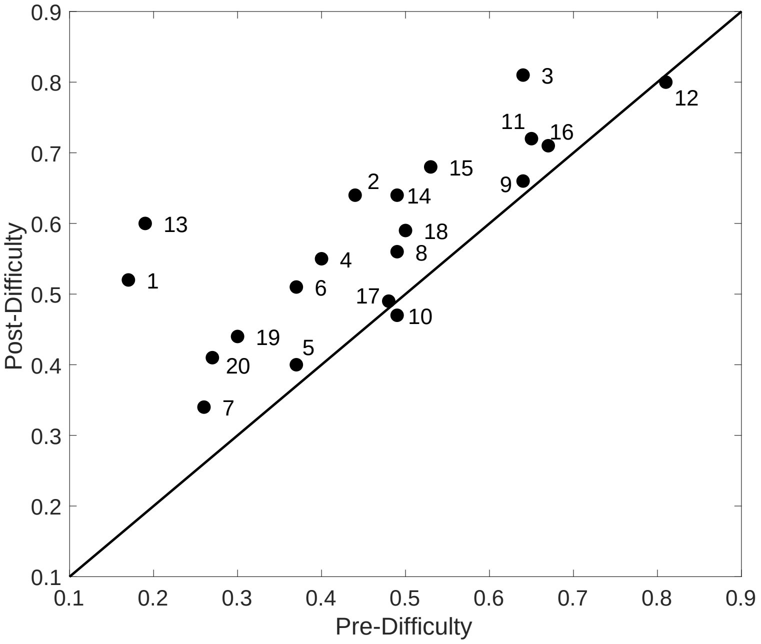


Figure 1. Pre-instruction item difficulty versus post-instruction item difficulty. Ideally, item difficulty values should be higher (a greater proportion of students answered the item correctly) post-instruction.

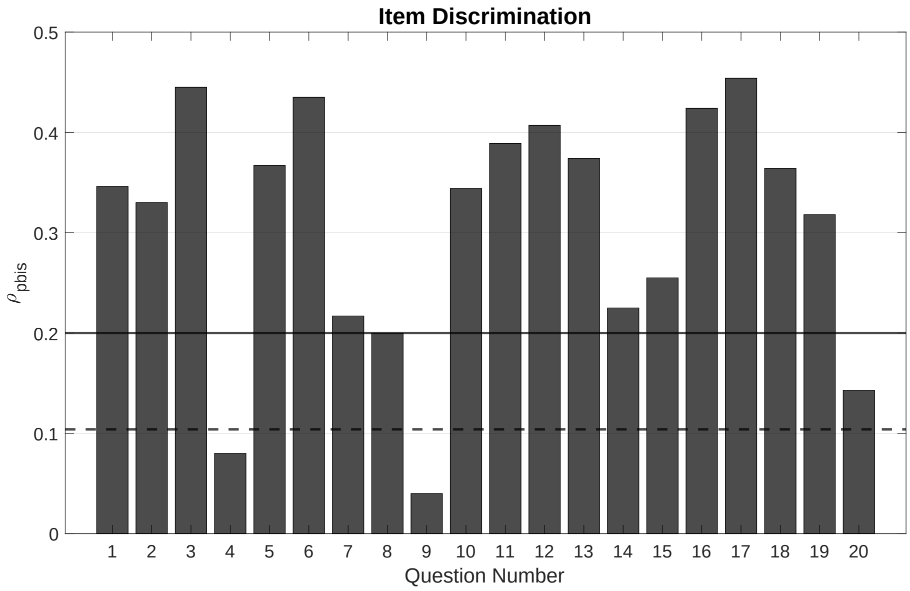


Figure 2. Post-instruction item discrimination (ρ_{pbis}) values. ρ_{pbis} values between 0.2-0.7 fall in the ideal range, while values greater than 0.104 are acceptable. The horizontal black line is at 0.2, while the dashed line is at 0.104.

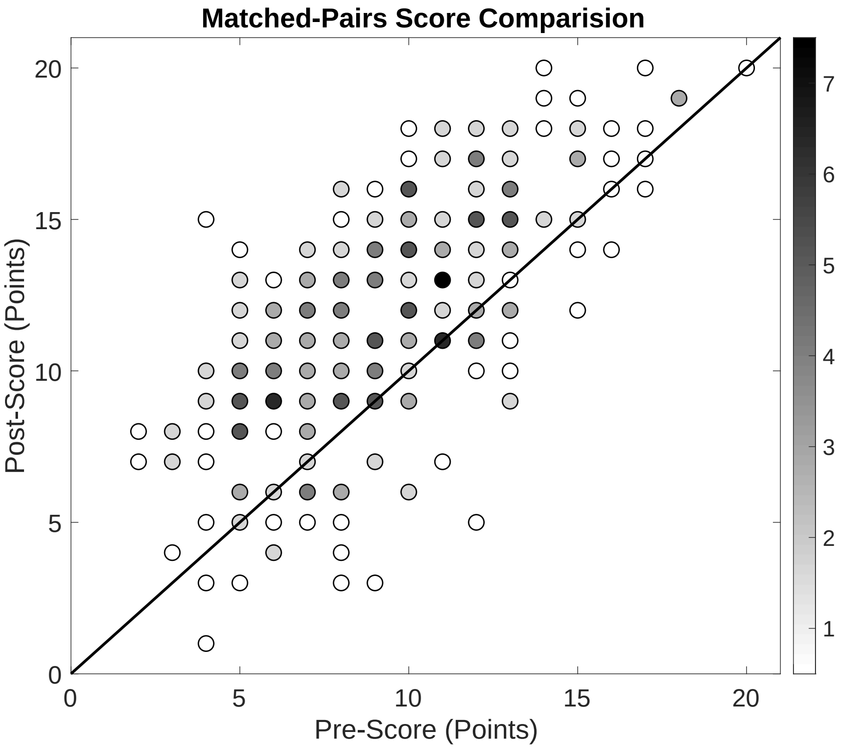


Figure 3. The pre-test scores versus post-test scores for the 287 students with matched pairs data. The diagonal line indicates the region of equal pre- and post-test scores. Students above the diagonal line performed better on the PFCI after instruction. The number of students represented by each datapoint is denoted on the colorbar (up to seven students).

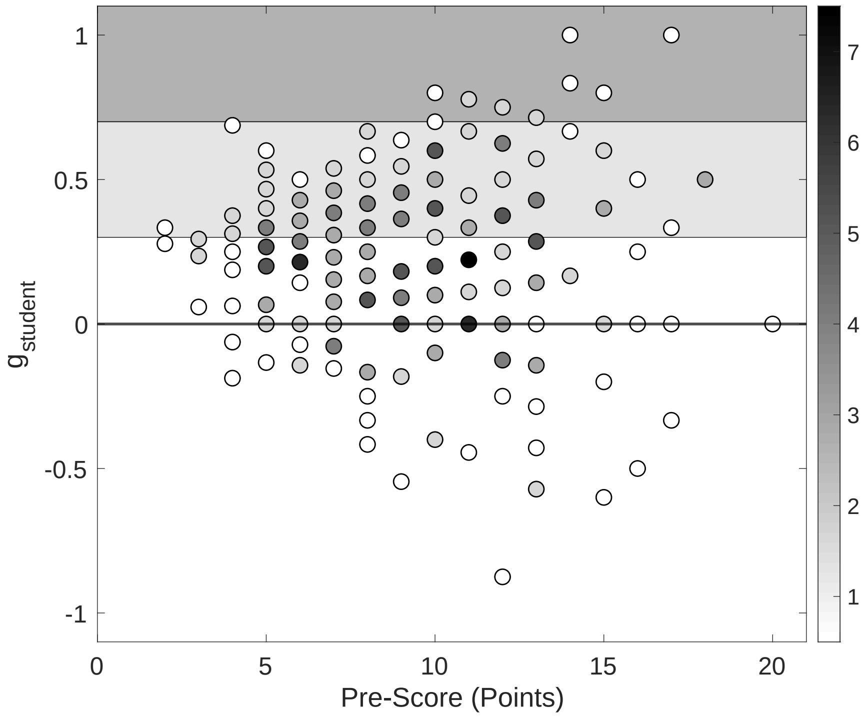


Figure 4. The pre-test scores versus normalized learning gain scores ($g_{student}$) for the 287 students with matched pairs data. The horizontal line represents zero normalized gain. The number of students represented by each datapoint is denoted on the colorbar (up to seven students). The regions corresponding to medium gain (0.3 - 0.7) and high gain (> 0.7) are shaded in light gray and darker gray, respectively.

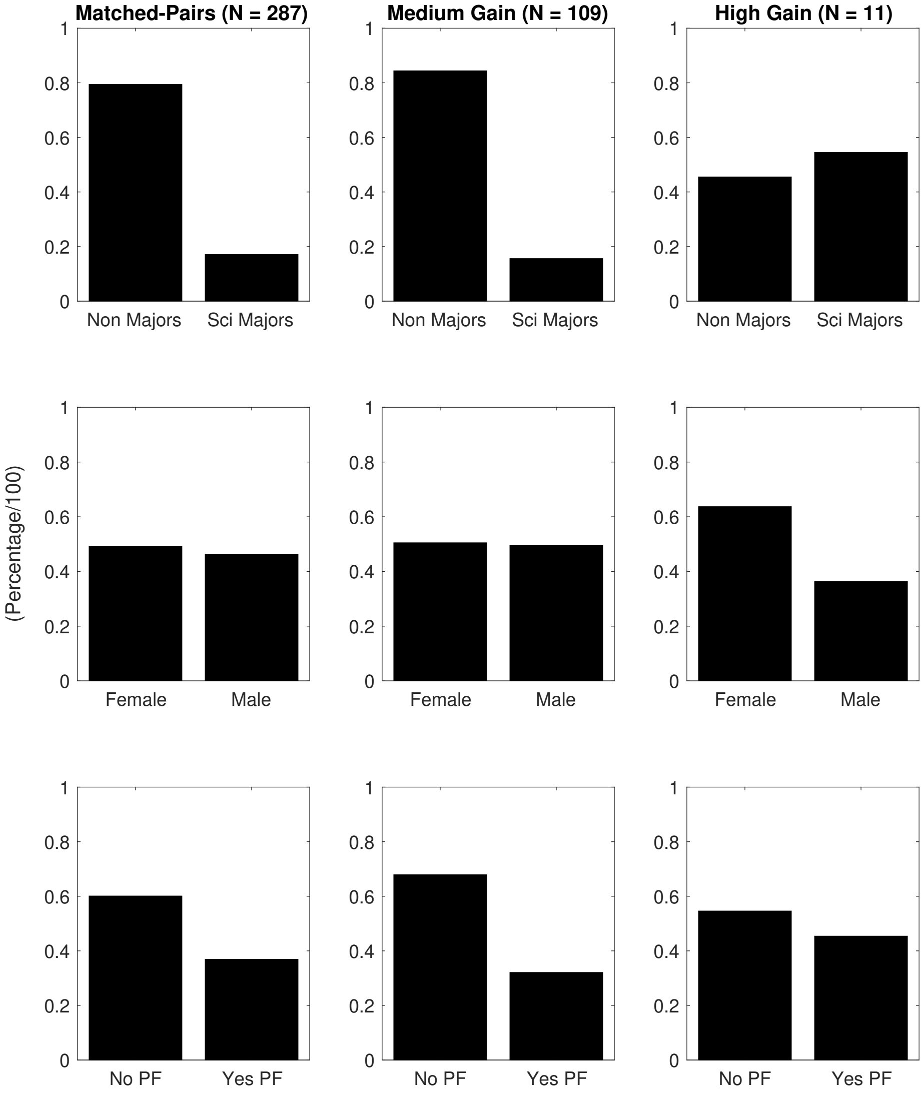


Figure 5. The demographic breakdown for the entire matched-pairs sample, the sample of students whose normalized gain scores were classified as 'medium gain', and the sample of students whose gain scores were classified as 'high gain'.

Fig 1. Pre-instruction item difficulty versus post-instruction item difficulty. Ideally, item difficulty values should be higher (a greater proportion of students answered the item correctly) post-instruction.

Fig 2. Post-instruction item discrimination (ρ_{pbis}) values. ρ_{pbis} values between 0.2-0.7 fall in the ideal range, while values greater than 0.104 are acceptable. The horizontal black line is at 0.2, while the dashed line is at 0.104.

Fig 3. The pre-test scores versus post-test scores for the 287 students with matched pairs data. The diagonal line indicates the region of equal pre- and post-test scores. Students above the diagonal line performed better on the PFCI after instruction. The number of students represented by each datapoint is denoted on the colorbar (up to seven students).

Fig 4. The pre-test scores versus normalized learning gain scores ($g_{student}$) for the 287 students with matched pairs data. The horizontal line represents zero normalized gain. The number of students represented by each datapoint is denoted on the colorbar (up to seven students). The regions corresponding to medium gain (0.3 - 0.7) and high gain (> 0.7) are shaded in light gray and darker gray, respectively.

Fig 5. The demographic breakdown for the entire matched-pairs sample, the sample of students whose normalized gain scores were classified as 'medium gain', and the sample of students whose gain scores were classified as 'high gain'.

Word Count: 7,149 words including tables, references, and figure captions (not the appendix)