Probing for Bias: Comparing Populations Using Item Response Curves

Paul J. Walter  
*St. Edward's University, pauljw@stedwards.edu*
Edward Nuhfer  
*California State University (retired), enuhfer@earthlink.net*
Crisel Suarez  
*Vanderbilt University, criselsuarez@gmail.com*

Follow this and additional works at: [https://scholarcommons.usf.edu/numeracy](https://scholarcommons.usf.edu/numeracy)

Part of the [Educational Assessment, Evaluation, and Research Commons](https://scholarcommons.usf.edu/numeracy), and the [Science and Mathematics Education Commons](https://scholarcommons.usf.edu/numeracy)

**Recommended Citation**
Walter, Paul J., Edward Nuhfer, and Crisel Suarez. "Probing for Bias: Comparing Populations Using Item Response Curves." *Numeracy* 14, Iss. 1 (2021): Article 2. DOI: [https://doi.org/10.5038/1936-4660.14.1.1357](https://doi.org/10.5038/1936-4660.14.1.1357)

Authors retain copyright of their material under a [Creative Commons Non-Commercial Attribution 4.0 License](https://creativecommons.org/licenses/by-nc/4.0/).
Probing for Bias: Comparing Populations Using Item Response Curves

Abstract

We introduce an approach for making a quantitative comparison of the item response curves (IRCs) of any two populations on a multiple-choice test instrument. In this study, we employ simulated and actual data. We apply our approach to a dataset of 12,187 participants on the 25-item Science Literacy Concept Inventory (SLCI), which includes ample demographic data of the participants. Prior comparisons of the IRCs of different populations addressed only two populations and were made by visual inspection. Our approach allows for quickly comparing the IRCs for many pairs of populations to identify those items where substantial differences exist. For each item, we compute the IRC dot product, a number between 0 and 1 for which a value of 1 occurs when the IRCs of the two populations are identical. We then determine whether the value of the IRC dot product is indicative of significant differences in populations of real students. Through this process, we can quickly discover bias across demographic groups. As a case example, we apply our metric to illuminate four SLCI items that exhibit gender bias. We further found that gender bias was present for non-science majors on those items but not for science majors.

Keywords

Item response curve, science and mathematics education, assessment

Creative Commons License

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License

Cover Page Footnote

Paul Walter is an associate professor of physics at St. Edward’s University in Austin, TX. Edward Nuhfer served as Director of Faculty Development and Educational Assessment and tenured professor of geology at four universities. His research interests are in metacognitive self-assessment, the role of the affective domain, and curricular design for reflective, higher-level thinking. He continues actively in writing, research, and assessment. Crisel Suarez received a B.S. degree in mathematics from St. Edward’s University. She is currently in the Ph.D. program at Vanderbilt University studying solar physics.

This article is available in Numeracy: https://scholarcommons.usf.edu/numeracy/vol14/iss1/art2
Introduction

Increasingly, investigators who construct assessment instruments such as concept inventories are simultaneously collecting demographic information from the participants they assess. This demographic information is critical to instrument validation because it reveals the degree to which each item produces unequal scores between groups and whether inequality exists for reasons investigators can seek to explain. For example, men and women seem equally capable of learning and achieving understanding. If an item testing content competency exerts no binary gender bias, the mean scores of two sufficiently large populations of women and men on that item should not be significantly different. Ideally, between groups for which there is no expected reason for any difference in levels of competency, such as men, women, or different ethnicities, the scores yielded by the item should not show systematic differences (i.e., bias). In reality, items often do show some bias.

For an ideal multiple-choice item with one correct answer, three distractors, and no bias, the distribution of selection frequencies of the incorrect distractor choices should be the same across different demographic groups. However, the actual result may not accord with the ideal in these cases either.

Consider a multiple-choice assessment instrument with 20 items, each having one correct choice and three distractors. Consider the demographic information collected for each participant as categories of binary gender, ethnicity, and class rank. One can quickly look at just these categories, of course. However, we also should look for bias between the sub-categories for each. These include ethnicity by gender, ethnicity by class rank and gender, and how each of these groups opted to select a distribution of distractors. It is easy to see that the number of cases to analyze and compare offers an intimidating data analysis task.

The inspection and discovery of the most critical biases are essential. Otherwise, starting by utilizing detailed analyses on every possible case commits an unnecessary and overwhelming expenditure of time and labor. To avoid such waste, investigators need to quickly pinpoint the items with the most substantial bias between particular paired populations and further investigate just these cases. In this paper, we provide a way to avoid discovering the significant biases through the laborious (and currently prevalent) process of manual inspection.

The focus here is to introduce a quantitative metric for comparing item response curves (IRCs) of two populations. We employ a data set derived from 12,187 respondents, exclusively undergraduates who took the Science Literacy Concept Inventory (SLCI; Nuhfer et al. 2016a). Appendix A, available as supplemental on the journal’s webpage for this paper, contains this data set, the Python script used to perform the data analysis for this work, and additional figures not shown in the main manuscript. The SLCI consists of 25 multiple choice items that probe the understanding of science as a way of knowing, and the instrument
captures considerable demographic data. We analyze the items on the SLCI using item response curves (IRCs), which were first introduced by Morris et al. (2006; 2012). IRCs show the percentage of students selecting a particular answer choice versus the overall score. As an example, Figure 1 shows the IRCs for first-generation students and non-first-generation students for Item 22 on the SLCI. The y-axis shows the percentage of respondents that chose a particular answer choice (A, B, C, or D), and the x-axis shows the overall score ranging from 0 to 25. The item response curves are slightly different for each population, and the focus of this work is to provide a means of comparison.

**Figure 1.** Item response curves (IRCs) for Item 22 on the SLCI for first-generation and non-first-generation students. IRCs show the percentage of students selecting a particular answer choice for an item as a function of the overall score. In this case, D is the correct answer choice since it is chosen by all students that score 25/25. Answer choice C is serving as a distractor for this item since it is the most commonly selected wrong answer for a wide range of overall scores. Most of the items on the SLCI do not have answer choices that serve as distractors. The IRCs for overall scores < 4/25 are not shown since so few students had such low scores.

Further, when comparing a pair of populations, there will be a plot similar to Figure 1 for each of the 25 items on the SLCI. When comparing many pairs of populations, as we will in this work, the number of plots to analyze by visual inspection quickly becomes overwhelming. In order to analyze the IRCs of different populations, we introduce a metric that compares the IRCs of two populations for an item and computes a single number to express their similarity or difference. The metric allows an investigator to probe for potential bias in any multiple-choice assessment, investigate whether different populations have the same misconceptions, and test the degree to which students of different populations (e.g., science and non-science majors) but with the same overall score find an item to have the same difficulty.
Literature Review

Options exist for analyzing multiple-choice tests and concept inventories. Ding and Beichner (2009) offer several approaches for performing data analysis on multiple-choice questions. They focus on five different methods: classical test theory, factor analysis, cluster analysis, item response theory (IRT), and model analysis. Given its close relationship to IRCs, we consider the similarities and differences of IRCs to IRT.

**Item Response Theory**

Researchers often employ IRT for analyzing the multiple-choice test items and comparing the performances of populations on concept inventories (Crocker and Algina 1986; Suen 2012; McDonald 2013; Baker and Kim 2017). IRT assumes there are latent traits (e.g., ability level) that provide a measure of the respondents’ understanding of a particular domain of content. The data input for each respondent consists of a 0 or 1 for each item to denote whether the answer was wrong or correct, respectively. For multiple-choice assessments, making the data binary (correct/incorrect) means we discard valuable information for the benefit of simplifying the analysis and interpretation (Tatsuoka 1983).

IRT models use logistic regression to fit parameters and generate item characteristic curves (ICCs) that show the probability of respondents selecting the correct answer as a function of their ability level. For any given item, the ICC is constructed by fitting to parameters (i.e., the discrimination, difficulty, and potentially a lower asymptote). The difficulty of an item is the value of the ability level that corresponds to an individual having a 50% likelihood of getting the item correct. A discrimination parameter derives from the maximum slope of the ICC at value of the difficulty for each item. Those items with the steeper ICC slopes at their difficulty values better discriminate between respondents of differing ability levels.

One way of testing an item for bias is to find signs of Differential Item Functioning (DIF), which are cases where the conditional probabilities for selecting the correct answer choice are different for one group than for another (Swaminathan and Rogers 1990). IRT provides a natural framework for DIF since IRT uses dichotomous scoring and the ICCs are monatomic functions of a latent trait (i.e., ability level). Potential signs of bias are indicated by the ICCs of two populations for an item not coinciding (Mellenbergh 1989; 2014). Appendix B of Watson et al. (2019) employed the 3PL IRT model to determine the difficulty and discrimination of each item on the SLCI and to account for the probability of random guessing (lower asymptote). Among other conclusions, they found that while some items are either more difficult for males or females, the SLCI was gender-neutral overall. We address this finding later under Discussion.
**Item Response Curves**

Morris et al. (2006) introduced IRCs analyzing data collected from respondents using the Force Concept Inventory (FCI; Hestenes et al. 1992). In contrast to IRT, which only considers whether responses are correct, IRCs track all possible responses (i.e., A–E on the FCI or A–D on the SLCI). Morris et al. (2006) made the simplifying assumption to use the overall score on the FCI as a substitute for ability level. Wang and Bao (2010) later showed a strong correlation between the two for the FCI. IRCs thus show the percentage of students selecting a particular answer choice versus the overall score. Morris et al. (2012) extended this work by using IRC analysis to examine each item on the FCI and determine how well each wrong answer choice functioned as a distractor for students of varied abilities.

Ishimoto et al. (2017) used IRC analysis to compare Japanese and American students’ performance on the Force and Motion Conceptual Evaluation (FMCE; Thornton and Sokoloff 1998). They found that the IRCs for the two populations were highly similar across most items. In those cases where differences appeared, they found that it was more likely due to contextual differences resulting from the translation than from a difference in the misconceptions held by American and Japanese students.

Our work introduces a metric to make a quantifiable comparison of the IRCs of two populations for any item on a concept inventory or multiple-choice assessment. Here we describe the metric and its utility. We neither rank our metric against other methods used to analyze data from multiple-choice assessments, nor do we divert from illustrating our metric’s utility by performing an exhaustive analysis of the SLCI.

**Data**

We collected our data in accordance with the approval to conduct human subjects research by IRB-105122 from 2010–2013 at CSU Channel Islands and IRB-13-019 from 2013–2021 at Humboldt State University to comply with all relevant federal guidelines and policies.

Here, we employ a data set consisting of 12,803 respondents to the 25-item Science Literacy Concept Inventory (SLCI). We removed 22 respondents who selected the same answer choice for all items. We also removed any respondents who either did not fully complete the demographic data or were not undergraduates, leaving our final data set containing the full responses and demographic data of 12,187 undergraduates. Numeracy has featured three other articles that used the SLCI (Nuhfer et al. 2016b; 2017; Watson et al. 2019). The psychometrics of this instrument were established in Nuhfer et al. (2016a), and the size of the database is adequate for our purpose in this paper. All 12,187 respondents entered the same
demographic information, including gender, ethnicity, whether English is a first language, class rank, and whether the student is a science major or has an interest.

Nuhfer et al. (2016b) emphasize the importance of investigators knowing what randomness looks like in any graphic that they intend to interpret, and they discuss cases in which researchers mistakenly interpreted graphical patterns generated from random noise as evidence of human test responses. In addition to the score distribution of the 12,187 students in our dataset (blue), Figure 2 also shows the score distribution of an equal number of simulated students who randomly guessed on every item.

![Figure 2](image)

**Figure 2.** The number of students with each score on the 25-item SLCI for 12,187 respondents. The mean is 17.7/25; the mode is 21/25; the median is 19/25. Also shown is the same number of simulated students who randomly guessed on each item.

**Methodology**

We begin an explanation of methodology by considering vectors and dot products. When one thinks of a vector, they normally think of a two- or three-dimensional arrow that points in some arbitrary direction, which would have two (i.e., $x$ and $y$) or three (i.e., $x$, $y$, and $z$) components, respectively. A vector $\mathbf{v}$ can be written as $\mathbf{v} = |\mathbf{v}| \hat{v}$ where $|\mathbf{v}|$ is the magnitude (e.g., length) of the vector, and the direction is given by the unit vector $\hat{v}$. The magnitude of a vector, which is a scalar, is given by

$$|\mathbf{v}| = \sqrt{\sum_i v_i v_i}$$
where the sum is over the components of the vector. For example, for a three-dimensional spatial vector
\[
|\mathbf{v}| = \sqrt{\sum_i v_i v_i} = v_x v_x + v_y v_y + v_z v_z.
\]

We can find a unit vector \( \hat{\mathbf{v}} \), which contains the direction information of the vector \( \mathbf{v} \), by dividing the vector \( \mathbf{v} \) by its magnitude.
\[
\hat{\mathbf{v}} = \frac{\mathbf{v}}{|\mathbf{v}|} = \frac{\mathbf{v}}{\sqrt{\sum_i v_i v_i}}
\]

Such a unit vector is normalized, meaning it has a magnitude of 1 and is unitless. A normalized vector satisfies the condition that \( \sum_i \hat{v}_i \hat{v}_i = 1 \).

A dot product of two vectors \( \mathbf{a} \) and \( \mathbf{b} \) is given by \( \mathbf{a} \cdot \mathbf{b} = \sum_i a_i b_i \). A dot product of two normalized vectors results in a scalar, a number, that signifies how well the vectors align with each other, i.e., the projection of one vector onto the other. For two normalized vectors that are at right angles to each other, the value of the dot product is 0. For two normalized vectors that are identical and thus point in the same direction, the value is 1. Figure 3 shows examples of the dot products of two normalized vectors. The values of a dot product can be negative when the vectors are anti-aligned. For our study, we will restrict ourselves to the case where the vectors’ components are either 0 or positive. Thus, the dot product values can range from 0 to 1. Vectors need not be three-dimensional, and can instead be \( N \)-dimensional, but two vectors must have the same dimension to be able to compute their dot product.

Figure 3. The dot products of two normalized vectors (meaning each vector has a length of 1) are shown. Left: The dot product of two vectors at right angles to each other is zero. Middle: The dot product of these two vectors is 0.5. Right: The dot product of a normalized vector with itself is 1.

**The IRC Dot Product**

Archibeque et al. (2018) measured the equity of conversations by members within each group by representing conversations with vectors and taking a dot product. We follow that example, and in our case, we are presented with a similar dilemma...
where there is a large amount of data (4 IRCs per item on the SLCI, 26 possible scores, 2 populations to compare) that we wish to boil down to a single number for each item on the SLCI (or for other common multiple-choice assessments) to allow easy comparisons of the IRCs of two populations. The following shows our process for computing what we will refer to as the IRC dot product, which compares how closely related the IRCs for two populations are for an item. We provide a Python script in Appendix A that performs the calculations outlined below for our SLCI dataset.

1. For each population, population a and population b, construct $a_{ijk}$ and $b_{ijk}$ where $i, j, k$ are indices; $i$ is the value of the answer choice (i.e., A–D on the SLCI), $j$ is the value of the overall score (i.e., 0–25 for the SLCI), and $k$ is the number of the item (i.e., 1–25 on the SLCI). Each component of $a_{ijk}$ is the number of respondents who had an overall score of $j$ that chose the answer choice $i$ for item $k$.

2. For each population, normalize over each answer choice. For instance, for population $a$

$$\hat{a}_{ijk} = \frac{a_{ijk}}{\sqrt{\sum_i a_{ijk} a_{ijk}}} \quad \forall j, k$$

and likewise, for population $b$.

3. Take the normalized dot products for two populations for each score and weight them by the number of students in each population for each score. The dot product of two populations for a particular score is given by $\sum_i \hat{a}_{ijk} \hat{b}_{ijk}$. In our case, this amounts to taking the dot product of 4-dimensional vectors since there are 4 possible answer choices. For a given item $k$, we will take this dot product for each possible score $j$. We also apply weighting since the number of students who have each score varies. More weight is given to the dot products for scores where both populations had relatively more students.

Thus, for our last step, we weight (by the number of students in each population with each particular score $j$) the normalized dot products for two populations by summing over the possible scores $j$. We thus introduce the IRC dot product

$$(\mathbf{a} \cdot \mathbf{b})_k \equiv \frac{\sum_j \sum_i n_{a,j} n_{b,j} \hat{a}_{ijk} \hat{b}_{ijk}}{\sum_j n_{a,j} n_{b,j}}$$

where $n_{a,j}$ and $n_{b,j}$ are the number of respondents who had score $j$ in population $a$ and $b$, respectively. Since all of the components of $\hat{a}_{ijk}$ and $\hat{b}_{ijk}$ are positive and given that $\hat{a}_{ijk}$ and $\hat{b}_{ijk}$ are each normalized, the values of $(\mathbf{a} \cdot \mathbf{b})_k$ for each of the $k$ items range from 0 (IRCs are very different) to 1 (IRCs are identical). For each item $k$, $(\mathbf{a} \cdot \mathbf{b})_k$ is a single number that is a measure of how close the IRCs of two populations are to each other.

In Step 2, we only normalize over the possible answer choices. This is because IRCs show the percentage of students for each population who select each answer choice as a function of their overall score, but they do not reflect how many students have each overall score. In Step 2, normalizing only over the answer choices
ensures that the dot products are independent of how many students in each population had a particular score \( j \). Thus, by normalizing over the answer choices and computing a dot product for each possible score \( j \), we are comparing the IRCs independent of the score distributions for each population.

To see why this matters, consider a simpler alternative to Equation 1, which is to take the dot products of normalized 104-component vectors (4 possible answer choices \( \times \) 26 possible scores), given by \( \sum_j \sum_i \hat{a}_{ijk} \hat{b}_{ijk} \). This formulation would also have values that range from 0 to 1. It also naturally builds in weighting since the components of \( \hat{a}_{ijk} \) and \( \hat{b}_{ijk} \) are the number of students who had a score \( j \) and selected answer choice \( i \) on item \( k \). The issue, however, is that each component of \( \hat{a}_{ijk} \) and \( \hat{b}_{ijk} \) depends upon the number of students who had score \( j \). For an item where the two populations have identical IRCs, the value of a dot product of normalized 104-component vectors need not be 1; it would also depend upon the shape of the score distributions for the populations. Only if the IRCs for an item were identical and the shape of the score distributions were identical would the dot product of normalized 104-component vectors have a value of 1.

In Equation 1, whenever the dot product of 4-component vectors \( \sum_i \hat{a}_{ijk} \hat{b}_{ijk} \) is 1 for each of the possible scores \( j \) on a particular item \( k \), which would mean that the IRCs of the two populations are identical, the weighting terms cancel and \((a \cdot b)_k = 1\). Thus, the IRC dot product, \((a \cdot b)_k\), in Equation 1 is the formulation that has the characteristics we are looking for in comparing only the IRCs while also ensuring that the scores with more students are weighted more heavily.

**Dot Product Ranges Resulting from Randomized Trials**

Our next goal is to ascertain whether the values of the IRC dot product, \((a \cdot b)_k\), given in Equation 1 are likely due to random chance that is reflective of the sample sizes or whether there are distinguishable differences between the IRCs of the two populations. For instance, we would like to know whether the value of the IRC dot product for an item provides an indication of whether an item is biased. To do so, we introduce a method for carrying out 10,000 randomized trials of computing \((a \cdot b)_k\) of simulated populations.

For each randomized trial, we first create two simulated samples of size \( N_a \) and \( N_b \), which are the total number of real students in populations \( a \) and \( b \), respectively. We apply weighted random sampling to assign a score \( j \) for each student of a given population that is weighted based on the distribution of scores for that population. Thus, the overall score distribution of our simulated students for a given population will closely resemble the score distribution of the real students of that population. As an example, when we compare the real male (\( N_{\text{male}} = 4,468 \)) and female (\( N_{\text{female}} = 7,719 \)) populations, the simulated male and female populations will have the same total number of simulated students in each
population and also have score distributions that are highly similar to that of the real male and female populations, respectively.

Next, for each simulated student, we assign an answer choice for each item. For the overall dataset ($N = 12,187$), let $P_{ijk}$ be the probability that students who have an overall score $j$ select answer choice $i$ on item $k$. We apply weighted random sampling for every simulated student, each of whom randomly selects an answer choice for each item that is weighted with the same probability $P_{ijk}$ as that of the overall population of real students who had a score $j$. This is repeated until we have $N_a$ simulated students making up population $a$ and $N_b$ simulated students making up population $b$ who all have been assigned an answer choice to each item as well as an overall score. The simulated populations have score distributions that mimic their respective populations and have selected answer choices for each item with the same probability as the overall population. We have constructed two simulated populations that are unbiased since they select answer choices from the same overall probability distribution.

Once we have our two simulated populations, we complete one randomized trial by following the procedure from the previous subsection to compute the IRC dot product, $(\mathbf{a} \cdot \mathbf{b})_k$, given in Equation 1. We repeat this process 10,000 times. For each item, the values of $(\mathbf{a} \cdot \mathbf{b})_k$ of the 10,000 randomized trials are spread out over a range of outcomes. From that overall range, we construct an interval that spans 95% of the outcomes (from 2.5% to 97.5%) of the $(\mathbf{a} \cdot \mathbf{b})_k$ values resulting from the randomized trials.

This 95% range is constructed based on the assumption that the item is unbiased since both simulated populations selected answer choices with the same probability $P_{ijk}$ of the overall population. If the value of the $(\mathbf{a} \cdot \mathbf{b})_k$ for the two real populations falls within this 95% range, then it is consistent with the hypothesis that the item is unbiased. When its value falls well below the 2.5% value (the low end of the range), then there is evidence that the deviation away from 1, a measure of how different the IRCs of the populations are, is too extreme to be the result of random chance for an unbiased item. Thus, in such cases, the item is either biased or there are clear differences between the two real populations.

**Results**

**Dot Product Approach Applied to SLCI Data**

Figure 4 shows the values of the IRC dot product, $(\mathbf{a} \cdot \mathbf{b})_k$, on the y-axis for first-generation students and non-first-generation students for each of the 25 items (x-axis) on the SLCI. In parentheses, the first number is the sample population (e.g., first-generation students), and the second number is the remaining population (e.g., non-first-generation students). For each item, we can assess whether the item shows more bias than can be explained by results of randomized trials of simulated
students. The values of the IRC dot product, \((a \cdot b)_k\), for first-generation and non-first-generation students, fall within the 95% range of the randomized trials for all but two items, and those two are not far outside of the range. Thus, with our data set, we do not detect substantially biased items when comparing the IRCs of first-generation and non-first-generation students.

![Figure 4](image)

**Figure 4.** The values of the IRC dot products, \((a \cdot b)_k\), of 4,546 first-generation undergraduate students and 7,641 non-first-generation undergraduate students for each item on the SLCI are shown in orange. The items with values for real students (orange dots) that fall within the 95% range of the randomized trials of unbiased simulated students (black error bars) are unbiased for our population of first-generation and non-first-generation students.

Figure 5 shows the values of IRC dot product for male and female students. Items 7, 15, 16, and 19 exhibit the most gender-bias. Based on IRT analysis, those items make up four of the five most difficult items (with Item 22 being the other) on the SLCI (Watson et al. 2019). For Items 7, 15, 19, the gender-bias is more pronounced than is observed for comparisons based on race and ethnicity, class rank, whether English is a first language, or whether the students are science majors. The plots of results for all other pairs of populations considered are available in Appendix A. When we compare the 95% ranges of the randomized trials for different items, we see a similar pattern in Figure 4 and Figure 5. Item 1 is the least difficult item, and it has the smallest range for the randomized trials and is also the closest to a value of 1. Item 15 is the most difficult item, and it has the largest range, and the top of its range (the 97.5 percentile of outcomes from the randomized trials) has the greatest departure from a value of 1. We will return to this in the Discussion.
Figure 5. The values of the IRC dot products (Eq. 1) for 4,468 male undergraduate students and 7,719 female undergraduate students for each item on the SLCI are shown in red. Items 7, 15, 16, and 19 show the most gender bias.

Figure 6 shows the values of the IRC dot product for science majors and non-science majors. Item 16 is biased between science majors and non-science majors, and it is gender-biased as well.

Figure 6. The values of the IRC dot products (Eq. 1) for 4,212 science majors and 7,975 non-science majors are shown in green. Item 16 is the only item that shows substantial bias.
Figure 7. The values of the IRC dot products (Eq. 1) for non-science majors where we compare genders are shown in green. Items 15 and 16 exhibit gender bias for non-science majors. As for Figure 5, Items 7, 15, 16, and 19 all show the most gender bias.

Figure 8. The values of the IRC dot products (Eq. 1) for science majors where we compare genders are shown in green. Unlike what was observed for non-science majors, Items 7, 15, 16, and 19 do not exhibit significant gender bias for science majors.

Given the large sample sizes involved, we can then probe further and consider comparing the IRCs of non-science majors based on gender (Figure 7) and science majors based on gender (Figure 8). For non-science majors, we find that Item 16 is gender-biased; however, Item 16 does not exhibit gender bias for science majors.
We also see that for Items 7, 15, and 19, the IRC dot products are similar for both non-science and science majors by gender. However, based on the location of the 95% range of randomized trials, the gender bias that is observed for non-science majors is not for science majors.

Figure 9 shows the values of the IRC dot product, \((\mathbf{a} \cdot \mathbf{b})_k\), for Pacific Islanders and non-Pacific Islanders. The y-axis has a different range from that of Figures 4–8, and the values of \((\mathbf{a} \cdot \mathbf{b})_k\) are considerably less than for the first-generation and gender comparisons shown in those figures. We expect this result since only 116 students in the sample identified as Pacific Islander, which are far too few to produce reliable IRCs. For such a small sample size, we would not expect the results for the real students to be different from random chance. The 95% range of the randomized trials confirms this expectation since the values of \((\mathbf{a} \cdot \mathbf{b})_k\) for the real students fall within the 95% range of the randomized trials of simulated students for all but one item, which itself is close to the randomized trial range. This result does not mean that we have shown that the items are unbiased toward Pacific Islanders, but only that it has not been observed for our sample. Making a determination of whether the items exhibit bias between Pacific Islanders and non-Pacific Islanders would require a larger sample size of Pacific Islanders. Further, for any items where we have not shown bias between two populations, the absence of evidence is not evidence of absence. We can only say that bias has not been observed for the item with the current sample, and that does not rule out observing bias for a different or larger sample of the two populations.

**Figure 9.** The values of the IRC dot products (Eq. 1) for 116 undergraduate students who identified as Pacific Islanders and 12,071 undergraduate students who identified as another race or ethnicity are shown in purple. The y-axis has a different range from that of Figures 4–8.
Discussion

Comparing the IRCs of two populations item-by-item by visual inspection alone can be overwhelming. In this work, we compared 17 pairs of populations (see Appendix A for those not shown in the main text), and each pair would require the visual inspection of the IRCs for each of the 25 items. Our approach enables comparing all of the IRCs for two populations for an item, such as those shown in Figure 1, and computing a single number using Equation 1, which is a metric for how close the IRCs are to each other. Visual inspection of the differences in the IRCs between two populations still has an important role but can be reserved for those items that exhibit bias for particular population pairs.

Discovering the bias of an item using our method does not tell us how the item is biased beyond identifying the two populations involved. A starting point for further investigation is then to examine the IRCs of the most biased items. Item 15, the most gender-biased item, asks students to demonstrate their ability to distinguish between the ways of knowing that typify either science or technology. Figure 10 shows the IRCs for male and female students for that item. For the overall score range from 10/25–22/25, males select the right answer choice more often than females. There are no apparent signs of gender bias from the wording in the item, but of all items on the SLCI, Item 15 exhibits the greatest mean difference in scores by gender. In this study, we do not attempt to explain why the observed differences exist, but note that we did not observe such a gender difference in the other SLCI item (Item 22) that also addresses distinctions between technology and science.

Figure 10. IRCs are shown for Item 15 on the SLCI for females and males. Item 15 is the most gender-biased on the SLCI. Note the difference in selecting the correct answer choice B (in blue) for the overall score range of 10/25–22/25.
Supplementary Figure 4 of Nuhfer et al. (2016a) draws a similar conclusion of Item 15 being the most biased toward males and Items 7, 15, 16, and 19 are the four most gender-biased items (all favoring males). However, they showed that the other items on the SLCI are biased toward females such that when one considers the SLCI as a whole, the gender bias cancels out, and the instrument overall is gender-neutral. Nuhfer et al. (2018) describe trying nine different items to address the ability to distinguish technology from science when developing the SLCI. Through all, the ability to distinguish between science and technology remained the most difficult to address by both experts and novices.

Item 16 asks students to identify which of four cases is not an example of modeling being used in science. We found Item 16 was gender-biased for non-science majors but not for science majors. Interestingly enough, the answer choice that serves as a distractor is the one that, based on its computational nature, most clearly is an example of what one considers to be modeling applied to science. So it may be that students are commonly missing the word “not” even though it is in all capital letters, and the distractor is not a misconception in this case, but instead is a sign of misreading the question. This does not explain why gender bias exists for non-science majors, which could perhaps be further elucidated with student interviews. From the IRCs (not shown, available in Appendix A), science majors found the item to be less difficult than non-science majors. The routine exposure science majors have of modeling being applied to science may explain the lack of (or a lesser amount of) gender bias.

The range of outcomes of the randomized trials by simulated students provides an effective way of detecting whether an item may be biased. When sample sizes are small, the 95% range of the simulated populations’ randomized trials is typically great enough such that the value of the IRC dot product for the real students falls inside it. Thus, we are unable to determine whether the item is biased for those populations. As was the case for when we compared the IRCs of Pacific Islanders with non-Pacific Islanders, having smaller IRC dot product values may be explained by the random noise one would expect for small sample sizes.

The randomized trial ranges were greater for the most difficult items, such as Item 15. Further, for those items, the 97.5 percentile (the top of the range) of outcomes of the randomized trials had its smallest values (i.e., the largest deviations from a value of 1). Since the simulated students who had a particular score are assigned an answer choice with the same probability as the total population of real students who had that score, having a greater than 95% range of outcomes for the randomized trials is likely a sign of relatively more of the real students being left to make an educated guess on an item. High-frequency oscillations in the IRCs, particularly in the intermediate score range containing most of the students, could be a sign of guessing. Such oscillations are apparent in the male and female science major IRCs, whose IRCs are noisier (they also have smaller sample sizes) than the
IRCs of male and female non-science majors. Those IRCs are not shown here but are available in Appendix A.

Conclusion

We have introduced a metric, referred to as the IRC dot product (Eq. 1), that enables the quantitative comparison of two populations’ IRCs for a single item on a concept inventory. The IRC dot product values range from 0–1, where a value of 1 means that the IRCs of two populations are identical for a specific item. Our approach not only allows for a quantitative comparison of the IRCs of a pair of populations, but also it provides a way of determining whether the results are a sign of bias in the item between the two populations. Bias in an item can be detected by comparing the IRC dot product values to those of a range of outcomes of randomized trials involving simulated students. Rather than comparing the IRCs for each item by visual inspection, our approach allows for processing large datasets and quickly comparing the IRCs of many pairs of populations to identify which items exhibit bias, and for which pairs of populations. After detecting item bias or differences between the IRCs that cannot be explained by random chance, an item-by-item investigation can focus on explaining possible causes for the observed differences.

In our example, used here as an illustrated application, our metric flagged the four items on the Science Literacy Concept Inventory that exhibited signs of gender bias. We were further able to show that the gender bias for those items was present for non-science majors but not science majors.

For items that have answer choices that function as distractors, the IRC dot product can also be used to determine whether different populations have the same misconceptions. We next hope to use our metric to examine data collected from the Force Concept Inventory (FCI). In the FCI, the items contain distractors that studies have shown signify a participant’s holding of particularly common misconceptions.

Acknowledgments

The authors would like to thank the faculty and students at numerous universities that participated in administering and taking the SLCI and thus provided the data set used in this work. We also thank the editor, Nathan D. Grawe, for his helpful suggestions.

References

Archibeque, Ben, Mary-Bridget Kustusch, Florian Genz, Scott Franklin, and Eleanor C. Sayre. 2018. “Qualitative Measures of Equity in Small Groups.”
Baker, Frank B., and Seock-Ho Kim. 2017. *The Basics of Item Response Theory Using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8.

Crocker, Linda, and James Algina. 1986. *Introduction to Classical and Modern Test Theory*. ERIC.

Ding, Lin, and Robert Beichner. 2009. “Approaches to Data Analysis of Multiple-Choice Questions.” *Physical Review Special Topics - Physics Education Research*, 5 (2): 020103. https://doi.org/10.1103/PhysRevSTPER.5.020103.

Hestenes, David, Malcolm Wells, and Gregg Swackhamer. 1992. “Force Concept Inventory.” *The Physics Teacher*, 30 (3): 141–58. https://doi.org/10.1119/1.2343497.

Ishimoto, Michi, Glen Davenport, and Michael C. Wittmann. 2017. “Use of Item Response Curves of the Force and Motion Conceptual Evaluation to Compare Japanese and American Students’ Views on Force and Motion.” *Physical Review Physics Education Research*, 13 (2): 020135. https://doi.org/10.1103/PhysRevPhysEducRes.13.020135.

McDonald, Roderick P. 2013. *Test Theory: A Unified Treatment*. Psychology Press. https://doi.org/10.4324/9781410601087.

Mellenbergh, Gideon J. 2014. “Item Bias Detection: Modern Approaches.” In *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06390.

Mellenbergh, Gideon J. 1989. “Item Bias and Item Response Theory.” *International Journal of Educational Research* 13 (2): 127–43. https://doi.org/10.1016/0020-7381(89)90002-5.

Morris, Gary A., Lee Branum-Martin, Nathan Harshman, Stephen D. Baker, Eric Mazur, Suvendra Dutta, Taha Mzoughi, and Veronica McCauley. 2006. “Testing the Test: Item Response Curves and Test Quality.” *American Journal of Physics*, 74 (5): 449–53. https://doi.org/10.1119/1.2174053.

Morris, Gary A., Nathan Harshman, Lee Branum-Martin, Eric Mazur, Taha Mzoughi, and Stephen D. Baker. 2012. “An Item Response Curves Analysis of the Force Concept Inventory.” *American Journal of Physics* 80 (9): 825–31. https://doi.org/10.1119/1.4731618.

Nuhfer, Edward B., Christopher B. Cogan, Carl Kloock, Gregory G. Wood, Anya Goodman, Natalie Zayas Delgado, and Christopher W. Wheeler. 2016a. “Using a Concept Inventory to Assess the Reasoning Component of Citizen-Level Science Literacy: Results from a 17,000-Student Study.” *Journal of Microbiology & Biology Education*, 17 (1): 143–55. https://doi.org/10.1128/jmbe.v17i1.1036.

Nuhfer, Edward, Christopher Cogan, Steven Fleischer, Eric Gaze, and Karl Wirth. 2016b. “Random Number Simulations Reveal How Random Noise Affects
the Measurements and Graphical Portrayals of Self-Assessed Competency.” *Numeracy*, 9 (1). https://doi.org/10.5038/1936-4660.9.1.4.

Nuhfer, Edward, Christopher Cogan, and Dawn Wright. 2018. “Unpacking the ‘S’ and ‘T’ of STEM: Overlooked Conceptual Differences Are Important.” In *Proceedings of the Pacific Division of AAAS, 17(1) 99th Annual Meetings Program with Abstracts*, 148.

Nuhfer, Edward, Steven Fleischer, Christopher Cogan, Karl Wirth, and Eric Gaze. 2017. “How Random Noise and a Graphical Convention Subverted Behavioral Scientists’ Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives.” *Numeracy*, 10 (1). https://doi.org/10.5038/1936-4660.10.1.4.

Suen, Hoi K. 2012. *Principles of Test Theories*. Routledge. https://doi.org/10.4324/9780203052211.

Swaminathan, Harihara, and H. Jane Rogers. 1990. “Detecting Differential Item Functioning Using Logistic Regression Procedures.” *Journal of Educational Measurement*, 27 (4): 361–70. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.

Tatsuoka, Kikumi K. 1983. “Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory.” *Journal of Educational Measurement*, 20 (4): 345–54. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.

Thornton, Ronald K., and David R. Sokoloff. 1998. “Assessing Student Learning of Newton’s Laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula.” *American Journal of Physics*, 66 (4): 338–52. https://doi.org/10.1119/1.18863.

Wang, Jing, and Lei Bao. 2010. “Analyzing Force Concept Inventory with Item Response Theory.” *American Journal of Physics*, 78 (10): 1064–70. https://doi.org/10.1119/1.3443565.

Watson, Rachel, Edward Nuhfer, Kali Moon, Steven Fleisher, Paul Walter, Karl Wirth, Christopher Cogan, Ami Wangeline, and Eric Gaze. 2019. “Paired Measures of Competence and Confidence Illuminate Impacts of Privilege on College Students.” *Numeracy*, 12 (2). https://doi.org/10.5038/1936-4660.12.2.2.