Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories

Jayson M. Nissen,[1,*] Robert M. Talbot,[2] Amreen Nasim Thompson,[2] and Ben Van Dusen[1]
[1]*Department of Science Education, California State University Chico, Chico, California 95929, USA*
[2]*School of Education and Human Development, University of Colorado Denver, Denver, Colorado 80217, USA*

(Received 25 August 2017; published 27 March 2018)

Measuring student learning is a complicated but necessary task for understanding the effectiveness of instruction and issues of equity in college science, technology, engineering, and mathematics (STEM) courses. Our investigation focused on the implications on claims about student learning that result from choosing between one of two commonly used metrics for analyzing shifts in concept inventories. The metrics are normalized gain (**g**), which is the most common method used in physics education research and other discipline based education research fields, and Cohen's *d*, which is broadly used in education research and many other fields. Data for the analyses came from the Learning About STEM Student Outcomes (LASSO) database and included test scores from 4551 students on physics, chemistry, biology, and math concept inventories from 89 courses at 17 institutions from across the United States. We compared the two metrics across all the concept inventories. The results showed that the two metrics lead to different inferences about student learning and equity due to the finding that **g** is biased in favor of high pretest populations. We discuss recommendations for the analysis and reporting of findings on student learning data.

DOI: 10.1103/PhysRevPhysEducRes.14.010115

I. INTRODUCTION

The methods for measuring change or growth and interpretations of results have been hotly discussed in the research literature for over 50 years [1]. Indeed, the idea of simply measuring a single state (let alone change) in an individual's understanding of a concept, conceptualized as a latent construct, is wrought with issues both philosophical and statistical [2]. Despite these unresolved issues, education researchers use measurement of growth for quantifying the effectiveness of interventions, treatments, and innovations in teaching and learning. Gain scores and change metrics, often referenced against normative or control data, serve as a strong basis for judging the efficacy of innovations. As researchers commonly measure change and report gains and effects, it is incumbent on researchers to do so in the most accurate and informative manner possible.

In this work, we collected data at scale and compared it to existing normative data to examine several statistical issues related to characterizing change or gain in student understanding. The focus of our analyses in this investigation are on student scores on science concept inventories (CIs). CIs are research-based instruments that target common student ideas or prior conceptions. These instruments are most often constrained response (multiple choice) and include these common ideas as attractive distractors from the correct response. There exist a multitude of CIs in use across biology, chemistry, and physics (our target disciplines) and in other fields (e.g., engineering and math). While CIs are common, the strength of their validity arguments varies widely and some lack normative data. All the CIs used in our work have at least some published research to support their validity and they align with our proposed uses for the scores.

A principal tool in quantitative research is comparison, which leads to the frequent need to examine different instruments and contexts. Complications arise in these cross contextual comparisons because the instruments used may have different scales and the scores may greatly vary between populations. For example, some CIs are designed to measure learning across one semester while others are designed to measure learning across several years. Instructors could use both instruments in the same course but they would, by design, give very different results. To compare the changes on the two instruments, researchers need to standardize the change in the scores, which researchers commonly do by dividing the change by a standardizing coefficient. Unlike in physics education research (PER) and other discipline based education research (DBER) fields, the social science and education research fields typically use a standardizing coefficient that is a measure of the variance of the scores.

*Corresponding author.
jnissen1@csuchico.edu

The most common gain measurement used in PER and other DBER fields when analyzing CI data is the average normalized gain, **g**, shown in Eq. (1) [3]. In this equation, the standardizing coefficient is the maximum change that could occur. Hake adopted this standardizing coefficient because it accounted for the smaller shift in means that could occur in courses with higher pretest means. He argued that **g** was a suitable measure because it was not correlated with the pretest mean, whereas the post-test mean and absolute gain were correlated with pretest mean and were not suitable measures. He also argued that this normalization allowed for "a consistent analysis over diverse student populations with widely varying initial knowledge states" [3] (p. 66) because the courses with lecture-based instruction *all* had low **g**, courses with active-engagement instruction primarily had medium **g**, and no courses had high **g**. Hake then used this reasoning to define high **g** ($\mathbf{g} > 0.7$), medium **g** ($0.7 > \mathbf{g} > 0.3$), and low **g** courses ($\mathbf{g} < 0.3$).

$$\mathbf{g} = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{100\% - \bar{x}_{\text{pre}}}. \tag{1}$$

Since Hake published his 1998 paper using **g**, it has been widely used, with the article being cited over 3800 times as this manuscript is being prepared as tracked by Google Scholar. Accordingly, there exists a large amount of gain data expressed in terms of **g** in the research literature, which serves as normative data for other studies. While the use of **g** does not align with the practices of the broader social science fields, it would be naive to dismiss this metric as unimportant. However, as noted above and discussed in more detail below, some issues exist with **g**.

The broader field of education research primarily uses the effect size metric as the preferred method for measuring change. The most commonly used effect size metric is Cohen's *d* [4]. In effect size metrics, a measure of the variance in the distribution of scores is the standardizing coefficient rather than the maximum possible gain, which **g** uses. An example of Cohen's *d* is given by Eq. (2), where the standardizing coefficient *s* is the pooled standard deviation of the pre- and post-tests (discussed further below).

$$d = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{s}. \tag{2}$$

Researchers have extensively investigated the utility and limitations of *d*, while the research investigating **g** is limited. In contrast to Hake's [3] earlier finding, Coletta and Phillips [5] found that **g** was correlated with pretest means. Willoughby and Metz [6] found that inferences based on **g** suggested gender inequities existed in college science, technology, engineering, and mathematics (STEM) courses even though several other measures indicated that there were no gender inequities in those courses. Furthermore, researchers use several different methods for calculating **g**, which can lead to discrepant findings [7,8]. Researchers have also identified issues for *d*. For example, *d* exaggerates the size of effects when measuring changes in small samples [9]. Cohen's *d* is based on the *t* statistic and the assumptions of normality and homoscedasticity in the test scores used to generate it [9]. CI data frequently fail to meet the assumptions of normality and homoscedasticity because of floor and ceiling effects and outliers. We expect that any problems that this creates for *d* are also applicable to **g**. However, we are not aware of any research on these assumptions pertaining to **g**.

II. PURPOSE

Both *d* and **g** have limitations. Our purpose in this investigation was to empirically compare concept inventory gains using both **g** and *d* to investigate the extent to which they lead to different inferences about student learning. In particular, our concern was that **g** favors high pretest populations, which leads to skewed measures of student learning and equity. This particularly concerned us because researchers use **g** as the de facto measure of student learning in PER and other DBER researchers have used it despite there being few investigations of the validity of **g** and known problems with its efficacy. We compared **g** to *d* since *d* is gaining use in DBER and is the comparable de facto measure in the much larger fields of sociology, psychology, and education research where researchers have extensively studied its validity, utility, and limitations.

III. BACKGROUND ON MEASURING CHANGE

In this section, we provide a foundation for our motivations and work. First, we discuss the development and use of CIs in undergraduate science education research. We then discuss statistical issues related to measuring change before reviewing the uses of the average normalized gain in analyzing scores from CIs. Finally, we discuss Cohen's *d* and its use in the context of best practices for presenting data and findings.

A. Rise in the use of CIs to measure student knowledge

CIs provide "data for evaluating and comparing the effectiveness of instruction at all levels" [10]. They typically consist of banks of multiple-choice items written to assess student understanding of canonical concepts in the sciences, mathematics, and engineering. Researchers generally develop CIs through an iterative process. They identify core concepts with expert feedback and use student interviews to identify common preconceptions and provide wording for distractors. CIs exist for core concepts in most STEM fields, see Ref. [11] for a thorough review and

discussion. Though it is unclear exactly how many CIs exist, one of the most widely used CIs, the Force Concept Inventory [10], has been cited more than 2900 times as this manuscript is being prepared as tracked by Google Scholar.

Researchers often use CIs as the outcome measures for evaluative studies to find out if an instructional intervention has an effect on learning relative to a control condition. To facilitate this use, researchers administer CIs pre- and post-instruction, and they compare gains observed in treatment groups to gains observed in a control condition. CIs tend to measure conceptual understanding at a *big picture* level. This means that if students conceive of science learning as a matter of primarily memorizing definitions and formulas (consistent with a more *traditional* conception of teaching and learning), they are unlikely to do well on most CIs. Several studies have used CIs to compare the impact of research-based pedagogies to more traditional pedagogies [3,12,13] and to investigate equity in STEM courses by comparing the knowledge and learning of majority and underrepresented minority students [14–17]. These types of investigations motivate instructors to adopt active learning in courses throughout the STEM disciplines [18].

Because researchers often compare scores for different CIs administered to different populations, they often use a change metric that is standardized and free from the original scale of the measurement. This change metric is often **g** for DBER studies, but some DBER studies have used d. One particular case that focused our current investigation on comparing **g** and d, was the use of **g** and test means to conclude "In the most interactively taught courses, the pre-instruction gender gap was gone by the end of the semester," [17] (p. 1). A finding that Rodriguez *et al.* [16] later called into question when their "analysis of effect sizes showed gender still impacted FCI scores and that the effect was never completely eliminated" [16] (p. 6).

B. Some issues in measuring change

Discussions in the measurement literature on quantifying change can be sobering. A classical and often cited work in this area is that of Cronbach and Furby [2], which raised issues of both the reliability and validity of gain scores. Based on classical test theory, they argue that the prime issue of reliability has to do with the systematic relationship between error components of true scores derived from independent, but "linked" observations. Consider a common situation in CI use in which the same test is given as both a pre- and post-test. One could argue that the observations (pre and post) are independent measurements since they are taken at different time points, but they are actually *linked* since the measurements are from the same group of students. Because those students had responded to the same instrument at the pretest administration, their post-test scores are likely correlated with their pretest scores due to a shared error component between the two scores. One can correct for this (often overstated) correlation due to the shared error components, but the correction is not always straightforward. Bereiter [1] calls this the "over correction under correction dilemma." Cronbach and Furby discuss this dilemma at length and offer various methods to dissattenuate the correlation. However, they seem to see these correction methods as a work around for the real issue of linked observations. In their summary discussion, Cronbach and Furby actually state that "investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways" (p. 80). Despite their persistent statistical issues, gain measurements are widely used in education research due to their great utility. Acknowledging these issues while leveraging the utility requires researchers to be diligent and transparent in their methods and presentation.

Another issue with gain scores has to do with the actual scale of the scores, which Bereiter refers to as the "physicalism-subjectivism dilemma." The issue here is related to the assumption of an interval scale on the construct of interest when using raw scores, or when using gain scores that are normalized on that same scale (as in using **g**). In other words, the gain metric (**g**) is scaled in terms of the measure itself (e.g., Newtonian thinking as measured by a CI) and is assumed to be measured on an interval scale. A potential solution here is to change the scaling to something that "seems to conform to some underlying psychological units," [1] (p. 5). In this case, the scaling factor (or "standardization coefficient") is not based on the scale of the measure (e.g., raw scores on a CI, or Newtonian thinking) but rather in a *standard* unit such as the variance of the score distribution. In this way, the gain metric is transformed out of the scale of the measure (e.g., Newtonian thinking) and into a construct independent, standardized scale (e.g., based on variance). Transforming the scale can make cross-scale comparisons possible, and also may highlight potential inequities brought on by remaining in the scale of the measure itself. This latter approach is how the dilemma is addressed when using the effect size metric (discussed further below). For a more detailed discussion of these issues related to measuring change in classical test theory, see Ref. [19].

C. The average normalized gain

Hake [3] developed the average normalized gain (**g**) as a way to normalize average gain scores in terms of how much gain could have been realized. Hake interpreted **g** from pre-post testing "as a rough measure of the effectiveness of a course in promoting conceptual understanding" (p. 5). His work was seminal in PER and led to the broad use of **g** throughout DBER. The breadth of its uptake led to at least three different methods for calculating **g** to be in common use. The original method proposed by Hake calculates **g** from the group means and is shown in Eq. (1). A second method that is more commonly used [13] is to calculate

the normalized gain for each individual student ($\mathbf{g}_I$) to characterize that student's growth, and to then average the normalized gains for all the individuals to calculate **g** for the group. Bao [7] provides an in depth discussion of the affordances of these two methods, but Bao and Hake both state that in almost all cases the two values are within 5% of one another.

Marx and Cummings [8] proposed a third method, normalized change (**c**), in response to several shortcomings of ($\mathbf{g}_I$). These shortcomings included a bias towards low pretest scores, a nonsymmetric range of scores ($-\infty$ to 1), and a value of $-\infty$ for any post-test score when the student achieves a perfect pretest score. These limitations inhibit the ability to calculate the average normalized gain for a class by averaging the individual student normalized gains. Instead, they offer a set of rules for determining **c** based on whether the student gains from pre- to post-test, worsens, or remains at the same score. Their metric results in values of **c** ranging from -1 to $+1$. However, **c** is still sensitive to the distribution of pre- and post-test scores in a way that might be "related to certain features of the population" [7], as it is still normalized on the same scale as the measure itself, an issue raised by Bereiter [1] and discussed above.

One particular concern with gain metrics, and with **g** specifically, has to do with the possibility that these metrics can be biased for or against different groups of students. As Rodriguez *et al.* [16] point out, researchers can define equity in several ways. This choice combined with potential bias in the gain metrics leaves open the possibility that results may not represent the actual status of equity in the classroom, for which gain metrics serve as a simple but incomplete indicator. For example, Willoughby and Metz [6] (p. 1) found that "males had higher learning gains than female students only when the normalized gain measure was utilized. No differences were found with any other measures, including other gain calculations, overall course grades, or individual exams." One might expect this to be the case when the pretest score is part of the standardization coefficient, since the pretest is likely correlated with previous education, and therefore opportunity and even socioeconomic status. Indeed, Coletta and Phillips [5] (p. 1) "found a significant, positive correlation between class average normalized FCI gains and class average preinstruction scores." This finding is aligned with Marx and Cummings [8] conclusion that **g** is biased by pretest scores, however, they found it was biased in the opposite direction.

D. The effect size metric

One of the most widely used standardized effect size metrics is Cohen's *d*. Cohen's *d* normalizes (i.e., scales) the difference in scores in terms of the standard deviation of the observed measurements. In essence, it is the difference between Z (standard) scores. This results in a "pure" number free from the original scale of measurement [4]. As a result, *d* meets the need for "… a measure of effect size that places different dependent variable measures on the same scale so that results from studies that use different measures can be compared or combined," Grissom and Kim [9].

As a consequence of using the standard deviation, *d* assumes that the populations being compared are normally distributed and have equal variances. Accordingly, the standard deviation used to calculate *d* is that of either sample from the population since they are assumed to be equal. However, in practical applications the pooled standard deviation, Eq. (3), of the two samples is used since the standard deviations of the two samples often differ. The pooled standard deviation (s_{pooled}) is a weighted average of the standard deviations of the two samples using the size of the samples (n) to weight the respective standard deviations. In the case of dependent data such as matched pre- and post-tests the sample size (n) for both samples is the same and can be factored out of the following:

$$s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}. \tag{3}$$

$$d_{\text{dep}} = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{\sqrt{s_{\text{pre}}^2 + s_{\text{post}}^2 - 2rs_{\text{pre}}s_{\text{post}}}} * \sqrt{2(1-r)}. \tag{4}$$

Using either the equal pre- and post-test standard deviations or the pooled standard deviations assumes that the samples (pre- and post-test) are independent and therefore does not take into account or correct for the correlation between measurements made at pre and post (the "dilemma" discussed above from Bereiter). The calculation for *d* accounting for the dependence between pre- and post-test [20] is shown in Eq. (4). Equation (4) is similar to Eq. (2) in that it represents the difference in the means divided by the standard deviation (in this case, s_{pooled}), noting that there are no sample sizes (n's) in Eq. (4) because they factor out of the equation since they are equal. Equation (4) differs in that it includes a correction factor to dissattenuate the effect size based on the correlation (r) between the pretests and post-tests. When the standard deviations are equal then the dependent and independent forms of Cohen's *d* are equal. If the two are not equal, then the dependent form of Cohen's *d* is always larger because the correlation accounts for some of the variance in the data and thereby reduces the standard deviation. Cohen's *d* can also be calculated from the *t*-test statistic and this can serve to further elucidate the dependent-independent issue. Dunlap *et al.* [21] present an example of calculating a *t* statistic between two means when assuming the samples are independent, and again when assuming dependence for the same sample means. When running dependent analyses the "…correlation between the measures reduces the standard error between the means, making the differences across the conditions more identifiable" [21] (p. 171).

Thus, taking into account the dependence between the data results in a larger t statistic because the difference in the means is divided by a smaller standard error. Cohen's d can be directly calculated from the dependent or independent t statistic. This is why the dependent form of d is always larger than the independent form. In practice, many researchers use the independent form of d given in Eq. (2) and do not account for correlations between the pre- and post-test. For example, Rodriguez *et al.* [16] used the dependent samples Cohen's d in their reanalysis of earlier pre-post data that did not provide the correlations between pretest and post-test scores. Dunlap *et al.* [21] recommend this practice, arguing that correlation does not change the size of the difference between the means but only makes the difference more noticeable by reducing the standard error. Morris and DeShon [22] agree that using the independent calculation for d with dependent data is an acceptable practice so long as all researchers are aware of the issue and any effect sizes being compared are calculated in the same way.

IV. RESEARCH QUESTIONS

Given our purpose of comparing $\mathbf{g}$ and d, our specific research questions were as follows:

(1) To what extent did the relationships between $\mathbf{g}$ and d and their relationships to the descriptive statistics used to calculate them indicate that they were biased toward different populations of students?

(2) To what extent did disagreements between $\mathbf{g}$ and d about the learning for groups of students with different pretest scores confirm any biases identified while investigating the first research question?

Based on previous research we expected differences in the degree to which d and $\mathbf{g}$ indicated that a phenomenon (e.g., learning gains or equity) was present. We expected the gain characterized by each metric to vary by student population due to differences in pretest scores across populations. This variation across pretest scores motivated our research because it could bias investigations of equity in college STEM learning. We used the second research question to test any biases we identified in a context (gender gaps) that is frequently investigated in the PER literature and to illustrate how bias in the measures used could skew the results of investigations.

V. METHODS

To answer these research questions, we used a large data set of student responses to nine different research-based CIs. This data set was large enough to provide useful and reliable comparisons of effect size measures and to represent CI data in general. We processed the data to remove spurious and unreliable data points and used multiple imputations (MI) to replace missing data. To simplify our analysis, we first investigated the similarity of the measures resulting from the three ways to calculate Hake's normalized gain, course averages ($\mathbf{g}$), averaged individual gains ($\mathbf{g_I}$), and normalized change ($\mathbf{c}$), to determine if they were similar enough we could conduct our further analyses using only one of those approaches. We then made several comparisons of the effect size measures for $\mathbf{g}$ and d to inform our research questions. We compared the effect size measures to one another and investigated the relationships of the effect size measures with pretest and post-test means and standard deviations to identify any potential biases in the effect size measures. To test any biases we found and to inform the effects of those biases, we compared the effect size measures for subpopulations within each course that have historically different pretest and post-test means.

A. Data collection and processing

Our general approach to data collection and processing was to collect the pre and post data with an online platform. We then applied filters to the data to remove pretests or post-tests that were spurious. Instead of only analyzing the data from students who provided both a pretest and a post-test, we used MI to include all the data in the analyses. Online data collection enabled collecting a large data set and filtering removed spurious and outlier data that was unreliable; using MI maximized the size of the sample analyzed and the statistical power of the analyses.

We used data from the Learning About STEM Student Outcomes (LASSO) platform that was collected as part of a project to assess the impact of learning assistants (LAs) on student learning [23,24]. LAs are talented undergraduates hired by university and two-year college faculty to help transform courses [25]. LASSO is a free platform hosted on the LA Alliance website [26] and allows faculty (LA using or not) to easily administer research-based concept inventories as pre- and post-tests to their students online. To use LASSO, faculty provide course-level information, select their assessment(s), and upload a list of student names and Email addresses. When faculty launch an assessment, their students receive Emails with unique links to complete their tests online at the beginning and end of instruction. Faculty can track students' participation and send reminder Emails. As part of completing the instrument, students answer a set of demographic questions. Faculty can download all of their students' responses and a summary report that includes a plot of their students' pre- and post-test scores and the course's normalized learning gains ($\mathbf{g}$), and effect sizes (Cohen's d).

We processed the data from the LASSO database to remove spurious data points and ensure that courses had sufficient data for reliable measurements. We filtered our data with a set of filters similar to those used by Adams *et al.* [27] to ensure that the data they used to validate the Colorado Learning Attitudes about Science Survey

(CLASS) were reliable. Their filters included number of items completed, duration of online surveys, and a filter question that directed participants to mark a specific answer. In our experience, Adams and colleagues discussion of filtering the data is unique for physics education researchers. Just as Von Korff *et al.* [13] found that few researchers explicitly state which **g** they used, we found that few researchers explicitly address how they filtered their data. For example, authors in several studies [28–30] that used the CLASS made no mention if they did or did not use the filter question to filter their data, nor do they discuss any other filters they may have applied. The lack of discussion of filtering in these three studies is not a unique choice by these authors. Rather, their choice represents the common practices in the physics education literature.

We included courses that had partial data for at least 10 participants to meet the need for a reliable measure of means without excluding small courses from our analyses. We removed spurious and unreliable data at the student and course level if any of the following conditions were met.

- A student took less than 5 minutes to complete that test. We reasoned that this was a minimum amount of time required to read and respond to the test questions.
- A student answered less than 80% of the questions on that test. We reasoned that these exams did not reflect student's actual knowledge.
- A student's absolute gain (post-test mean minus pretest mean) was 2 standard deviations below the mean absolute gain for that test. In these cases, we removed the post-test scores because we reasoned that it was improbable for students to complete a course and unlearn the material to that extent.
- A course had greater than 60% missing data. Low response rates may have indicated abnormal barriers to participating in the data collection that could have influenced the representativeness of the data from those courses.

Filter 1, taking less than 5 minutes, removed 364 students from the data set. Filter 2, completing less than 80% of the questions, removed 10 students from the data set. Filter 3, a negative absolute gain 2 standard deviations below the mean, removed 0 students but did remove 43 post-tests. Removing the courses with more than 60% missing data removed 27 courses and 1116 students from the analysis.

To address missing data, we performed multiple imputations (MI) with the Amelia II package in R [31]. The most common method for addressing missing data in PER is to use listwise deletion to only analyze the complete cases, discarding data from any student who did not provide both the pretest and post-test; though, we know of at least one study in PER that used MI [32]. We used MI because it has the same basic assumptions of listwise deletion but it reduces the rate of type I error by using all the available information to better account for missing data [33]. This leads to much better analytics than traditional methods such as listwise deletion [34] that, while they "…have provided relatively simple solutions, they likely have also contributed to biased statistical estimates and misleading or false findings of statistical significance," [35] (p. 400). Extensive research indicates that in almost all cases MI produces superior results to listwise deletion [36,37].

MI addresses missing data by (i) imputing the missing data m times to create m complete data sets, (ii) analyzing each data set independently, and (iii) combining the m results using standardized methods [37]. The purpose of MI is not to produce specific values for missing data but rather to use all the available data to produce valid statistical inferences [36].

Our MI model included variables for CI used, pretest, and post-test scores and durations, first time taking the course, and belonging to an underrepresented group for both race or ethnicity and for gender. The data collection platform (LASSO) provided complete data sets for the CI variables and the student demographics. As detailed in Table I, either the pretest score and duration or the post-test score and duration was missing for 42% of the students. To check if this rate of missing data was exceptional, we identified 23 studies published in the American Journal of Physics or Physical Review that used pre-post tests. Of these 23 studies, 4 reported sufficient information to calculate participation rates [28–30,38]. The rate of missing data in these 4 studies varied from 20% to 51% with an average of 37%. The 42% rate of missing data in this study was within the normal range for PER studies using pre-post tests.

Based on the 42% rate of missing data we conducted 42 imputations because this is a conservative number that will provide better results than a smaller number of imputations [39]. We analyzed all 42 imputed data sets and combined the results by averaging the test statistics (e.g., means, correlations, and regression coefficients) and using Rubin's rules to combine the standard errors for these test statistics [40]. Rubin's rules combines the standard errors for the analyses of the MI data sets using both the within-imputation variance and the between-imputation variance with a weighting factor for the number of imputations used. For readers interested in further seeking more information on MI, Schafer [40] and Manly and Wells [36] are useful overviews of MI. All assumptions were satisfactorily met for all analyses.

TABLE I. Data after each filter was applied.

	None	Time	Completion	Gain	≥60% missing
Courses	119	116	116	116	89
Students	6041	5677	5667	5667	4551
Pretests	5339	4922	4899	4899	3842
Post-tests	4204	3693	3685	3642	3335
Matched	3502	2973	2917	2874	2626

TABLE II. Correlations between the three forms of normalized gain for each course. *** indicates $p < 0.001$

	c	g
g	0.99***	
$\mathbf{g_I}$	0.93***	0.93***

B. Investigating the effect size measures

To identify and investigate differences between the effect size measures, we used correlations and multiple linear regressions (MLR) to investigate the relationships between the effect size measures and the test means and standard deviations. Correlations informed the variables we included in the MLR models.

We calculated Cohen's d for each course using the independent samples equation, Eq. (2). We used this measure because it is the most commonly used in the physics education research literature and because we expected it to have little to no impact on the analyses [22], which we discussed in Sec. III.D.

To test biases in the effect sizes and their effects on CI data, we used the male and female effect size measures in the aggregated data set. We separated these two groups because male students tend to have higher pretest and post-test means on science concept inventories than female students [14,41]. Thus, gender provided a straightforward method of forming populations with different test means and standard deviations. Gender also allowed us to frame our analysis in terms of equity of effects. We defined equity as being the case where a course does not increase preexisting group mean differences. This definition means that for a course to be equitable the effect on the lower pretest group is equal to or larger than the effect on the higher pretest group.

For this analysis we calculated the effect sizes for males and females separately. For each effect size measure we then calculated the difference between males and females effect sizes, for example, $\Delta_{\mathbf{g}} = \mathbf{g}_{\text{male}} - \mathbf{g}_{\text{female}}$. If males in a course had a larger effect size than females in the course then that course was inequitable, $\Delta_{\mathbf{g}} > 0$. This created four categories into which any two effect size measures would locate each course. Two categories for agreement where both effect sizes said it was either equitable or inequitable and two categories for disagreement where one said equity and the other said inequity. If one of the effect size measures was biased and indicated larger effects on higher pretest mean populations than we expected that one type of disagreement would occur more frequently than the other type. To easily identify differences in the number of courses in the disagreement categories and the size of those disagreements, we plotted the data on a scatter plot. We tested the statistical difference in the distributions using a chi square test of independence with categories for each effect size measure and whether they indicated equity or inequity.

C. Simplifying the analysis

The multiple methods for calculating normalized gain for a course complicated our purpose of comparing normalized gain and Cohen's d. Therefore, we compared normalized gain calculated using each of the three common methods, which are described in Sec. III. C, for each course. We calculated **g** using the average pretest and post-test scores for the course. We calculated the course $\mathbf{g_I}$ and **c** by averaging the individual student scores for each course. Correlations between all three measures were all large and statistically significant, as shown in Table II. The scatter plots for these three measures are shown in Fig. 1. These results indicated that all three measures were very similar. Therefore, we only used the normalized gain calculated using course averages, **g**, in our subsequent analyses.

The filters we applied to the data likely minimized the differences between $\mathbf{g_I}$ and the other two forms of normalized gain. As Marx and Cummings [8] point out, $\mathbf{g_I}$ is asymmetric. Students with high pretest scores can have very large negative values for $\mathbf{g_I}$, as low as approximately −32, but can only have positive values up to 1. We focused on filtering out spurious and unreliable data that

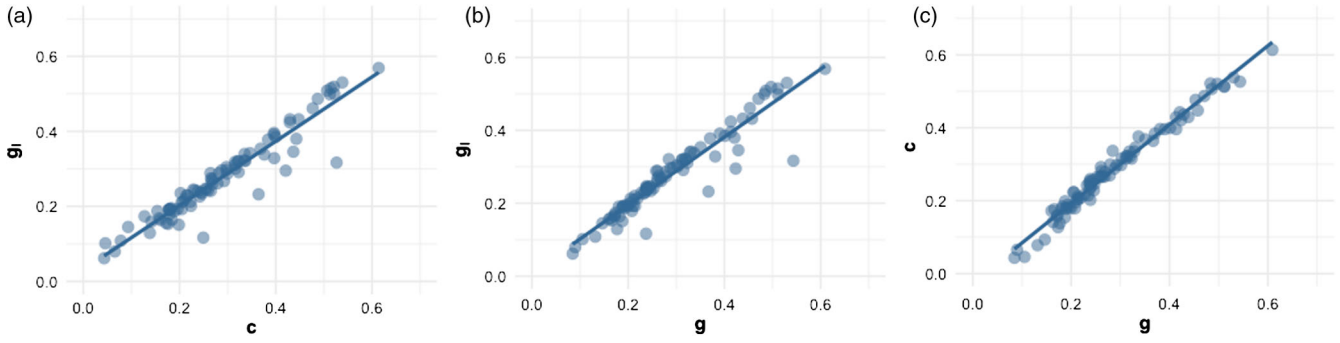

FIG. 1. Scatter plots comparing the course average value for the three forms of normalized gains.

TABLE III. Correlations between the effect size measures and test statistics, including pretest mean (Pre. Mean), pretest standard deviation (Pre. S.D.), post-test mean (Post. Mean) and post-test standard deviation (Post. S.D.). $*p < 0.05$. $**p < 0.01$. $***p < 0.001$.

	d	**g**	Pre. Mean	Post. Mean	Gain	Pre. S.D.
g	0.75***					
Pre. Mean	−0.11	0.43***				
Post. Mean	0.43***	0.87***	0.81***			
Gain	0.87***	0.93***	0.11	0.67***		
Pre. S.D.	−0.18	0.44***	0.67***	0.65***	0.24*	
Post. S.D.	−0.24	0.10	−0.01	0.08	0.15	0.42***

would have likely produced many large negative $\mathbf{g_I}$ values for individual students and resulted in larger differences between $\mathbf{g_I}$ and the other two normalized gain measures. Nonetheless, the notable differences between the three normalized gain metrics all occurred for $\mathbf{g_I}$ being much lower than the other two metrics.

VI. FINDINGS

A. Relationship between g and *d*

Investigating the relationship between **g** and d indicated that there was a large positive relationship between the two measures and it was statistically reliable: $r = 0.75$, $p < 0.001$. This indicated that d and **g** shared approximately half of their variance in common ($r^2 = 0.56$). Because these two measures serve the same purpose, the 44% that they *do not* have in common was a large amount. Further investigating the correlations between the effect sizes and their related descriptors, shown in Table III, revealed large differences between d and **g**. The correlations between d and both pretest mean and pretest standard deviation were small to very small and were not statistically reliable. In contrast, **g** was moderately to strongly correlated with both pretest mean and pretest standard deviation. These correlations between **g** and pretest statistics (0.43 and 0.44, respectively) indicated that approximately one-fifth of the variance in normalized gains was accounted for by the score distributions that students had prior to instruction. In contrast, d was only weakly associated with both pretest mean and pretest standard deviation. Three of these relationships are shown in Fig. 2. These relationships were strong evidence that **g** was positively biased in favor of populations with higher pretest scores.

To inform the size of this bias, we ran several models using MLR with **g** as the dependent variable and independent variables for d, pretest mean, and pretest standard deviation. We used **g** as the dependent variable because this was consistent with the correlations between **g** and pretest mean indicating that **g** was positively biased by pretest means, whereas correlations indicated that d was not biased. The linear equation for the final model is given in Eq. (5). Our focus in these MLRs was on the additional variance explained by each variable in the models, which we measured using the adjusted r^2. We did not focus on the coefficients β for each variable:

$$\mathbf{g} = \beta_0 + \beta_1 \times d + \beta_2 \times \text{Mean}_{\text{pre}} + \beta_3 \times \text{S.D.}_{\text{pre}} \quad (5)$$

The four models for the MLR are shown in Table IV. All the models were statistically significant ($p < 0.05$). Model 1 only included d and shows that d and **g** shared 55% of the same variance as indicated by the adjusted r^2 value. Adding either pretest mean or pretest standard deviation to the model markedly increased the explained variance to either 82% or 89%, model 2 and model 3, respectively. Including all three variables in model 4 explained 92% of the variance in **g**. We interpreted this as indicating that the disagreements between d and **g** were largely explained by the pretest means and standard deviations. Because the pretest means and standard deviations explained such a large proportion of the unexplained variance from model 1 and the

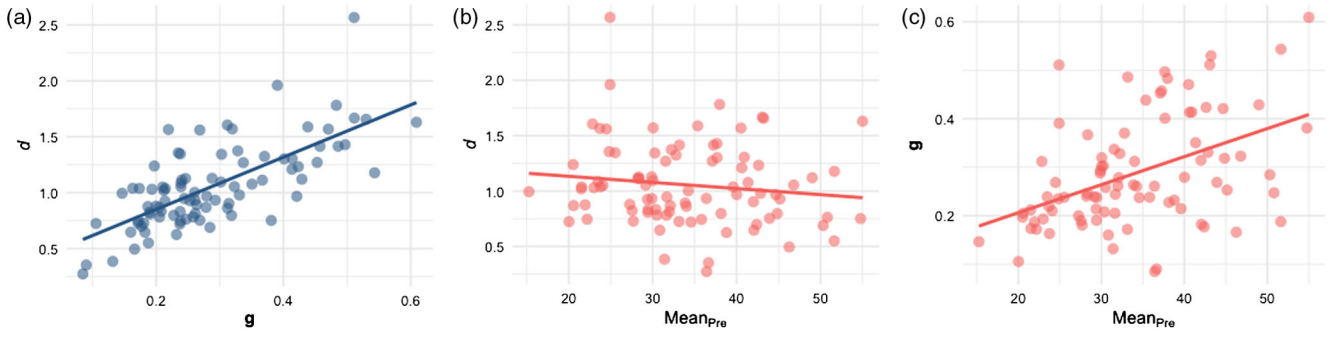

FIG. 2. Scatter plots for (a) d and **g**, (b) d and pretest mean, and (c) **g** and pretest mean.

TABLE IV. MLR exploring relationships between **g** and dependent variables (D.V.) for *d*, pretest mean (Pre. Mean), and pretest standard deviation (Pre. S.D.).

Model	Model 1		Model 2		Model 3		Model 4	
$r^2(\%)$	55.6		82.4		89.0		92.1	
adj. $r^2(\%)$	55.0		82.0		88.7		91.8	
p	<0.001		<0.001		<0.001		0.02	
D.V.	β	*p*	β	*p*	β	*p*	β	*p*
Intercept	0.05	0.05	−0.21	<0.001	−0.21	<0.001	−0.26	<0.001
d	0.22	<0.001	0.24	<0.001	0.25	<0.001	0.25	<0.001
Pre. Mean			0.01	<0.001			0.01	<0.001
Pre. S.D.					0.01	<0.001	0.01	<0.001

correlations indicated that pretest mean and pretest standard deviation were much more strongly related to **g** than to *d*, these results indicated that **g** was biased in favor of groups with higher pretest means.

B. Testing the bias in g using populations with different pretest scores

Results from the MLR model 2 indicated that a class's pretest mean explained 27% of the variance in a class's **g** value that was not explained by *d*. If **g** is biased in favor of high pretest groups, as the MLR and correlations indicated, then we expected the disagreements between **g** and *d* to skew such that they indicated a bias for **g** in favor of the high pretest population. To visualize potential bias in **g** we plotted the difference in *d* on the x axis and the difference in **g** on the y axis in Fig. 3. The course marker color shows whether male or female students' pretest means were higher. Almost all of the markers (41 of 43 courses) indicated that male students started with higher pretest means and that the data were consistent with our focus on equity being a larger effect on female students. In total, **g** showed a larger effect on males in 33 out of 43 courses, whereas *d* indicated a larger effect on males in 22 out of 43 courses. Figure 3 illustrates this bias in **g** in the difference between quadrants II and IV. A chi squared test of independences indicated that these differences were statistically reliable: $\chi^2(1) = 6.10$, $p = 0.013$. This difference confirmed that **g** was biased in favor of the male population, showing that **g** is biased in favor of populations with higher pretest means. This bias implies that **g** is not a sufficiently standardized change metric to allow comparisons across populations or instruments with different pretest means and is not a suitable measure of effects.

VII. DISCUSSION

To simplify our comparison of the statistical merits of using **g** and *d* to measure student learning, we first determined what differences there were between the three methods of calculating **g**. Our analysis showed that the three methods for calculating normalized gain scores were highly correlated ($r \geq 0.93$). The high level of correlation between the normalized gain values indicated that it made

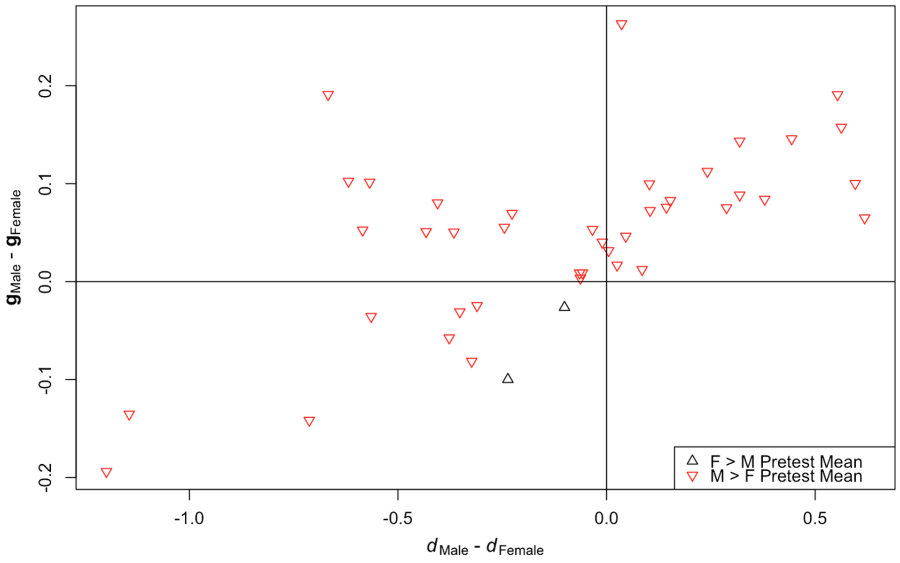

FIG. 3. Comparison of gender differences for *d* and **g**.

little difference which method we used. This result was encouraging given that many researcher report **g** scores without discussing which method of calculation they used [13]. The scatter plots (Fig. 1) for the three measures of **g** indicated that the large disagreements between the measures that occurred were cases in which $\mathbf{g_I}$ was much lower than both **g** and **c**. This discrepancy is consistent with the negative bias in possible $\mathbf{g_I}$ scores that led Marx and Cummings [8] to develop **c**. The filters we used to remove unreliable data likely increased the agreement we found between $\mathbf{g_I}$ and the other normalized gain measures. However, there were several courses where $\mathbf{g_I}$ was noticeably lower than **g** and **c**. These disagreements indicate two potential problems in the existing literature. Some studies using $\mathbf{g_I}$ may have underestimated the learning in the courses they investigated due to the oversized impact of a few large negative $\mathbf{g_I}$ values. Alternatively, some studies may have filtered out data with large negative $\mathbf{g_I}$ values but not explicitly stated this filtering occurred. Both situations are consistent with Von Korff and colleague's [13] statement that few researchers explicitly state which measure of normalized gain they used. Either situation or a combination of the two make it difficult for researchers to rely on and to replicate the work of those prior studies.

Our comparisons of **g** and d revealed several meaningful differences that indicated that **g** was biased in favor of high pretest populations. The correlation between **g** and d was strong ($r = 0.75$, $p < 0.001$) but was markedly smaller than the correlations between the three different methods of calculating **g** ($r \geq 0.93$). This correlation of 0.75 meant that **g** and d shared only 56% of their variance. MLRs indicated that pretest mean and standard deviation explained most of the difference between **g** and d; d, pretest mean, and pretest standard deviation accounted for 92% of the variance in **g**. Given that **g** was correlated with these pretest statistics much more strongly than d, we concluded that **g** is biased in favor of populations with high pretest means. We recommend that researchers avoid using all forms of normalized gain and instead report Cohen's d and the descriptive statistics used to calculate it, including the correlation between pretest and post-test scores.

This bias of **g** in favor of populations with high pretest means is problematic. The dependence of **g** on pretest privileges populations of students who come into a class with more disciplinary knowledge or who perform better on multiple choice exams. This bias disproportionately affects students from traditionally underrepresented backgrounds such as women in physics. When comparing the learning of males and females in our data set, **g** identified males as learning more in 33 of 43 courses (77%) while d only identified males as learning more in 23 of 43 courses (53%), nearly cutting the rate by 1/3 (Fig. 3). This difference in measurement indicated that **g** should not be used for investigations of equity as it overestimated student inequities. Researchers are better served by using statistical methods that analyze individual student's post-test scores while controlling for their pretest scores and other variables of interest. All researchers should ensure that they report sufficient descriptive statistics for their work to be included in meta-analyses.

VIII. CONCLUSION AND INFERENCES

The bias in **g** can harm efforts to improve teaching in college STEM courses by misrepresenting the efficacy of teaching practices across populations of students and across institutions. Students from traditionally underrepresented backgrounds are disproportionately likely to have lower pretest scores, putting them at a disadvantage when instructors make instructional or curricular decisions about an intervention's efficacy based on **g**. For example, **g** likely disadvantages instructors who use it to measure learning in courses (e.g., nonmajor courses) or are at institutions (e.g., 2-year colleges) that serve students who have lower pretest means. This is particularly important for faculty at teaching intensive institutions where evidence of student learning can be an important criterion for tenure and promotion.

Comparing the impact of interventions across settings and outcomes in terms of gain scores requires some form of normalization. Normalized learning gain (**g**) and Cohen's d both employ standardization coefficients to account for the inherent differences in the data. Hake developed **g** to account for classes with higher pretest means potentially having lower gains due to ceiling effects. By focusing on ceiling effects, **g** implicitly assumes that any population with a higher pretest score will have more difficulty in making gains than lower pretest populations. This assumption contradicts one of the most well-established relationships in education research that prior achievement is a strong predictor of future achievement. Thus, **g**'s adjustment for potential ceiling effects appears to overcorrect for the problem and results in **g** being biased in favor of populations with higher pretest means.

Using standard deviation as the standardization coefficient in Cohen's d helps to address ceiling effects in that measure. When ceiling effects occur the data compresses near the maximum score. This compression causes the standard deviation to decrease which increases the size of the d for the same raw gain. Cohen's d also corrects for floor effects by this same mechanism. Instruments that have floor and ceiling effects are not ideal for research because they break the assumption of equal variances on the pre- and post-tests and because they are poor measures for high or low achieving students. Instruments designed based on classical test theory, such as the CIs used in this study, mainly consist of items to discriminate between average students and have few items to discriminate between high-performing students or low-performing students. Cohen's d may mitigate the limitations of these instruments for measuring the learning of high or low pretest populations of students by accounting for the distribution of tests scores.

When the standard deviation is smaller, as with floor or ceiling effects, the probability of change is lower (i.e., learning is harder) so Cohen's d is larger in these cases for the same size change in the means.

In addition to reporting Cohen's d, researchers should include descriptive statistics to allow scholars to use their work in subsequent studies and meta-analyses. These descriptive statistics should include means, standard deviations, and sample sizes for each measure used, and correlations between the measures. We include correlations on this list because of the dependent nature of CI pre-post testing is not taken into account by the change metrics we have presented in this paper. As discussed in the background section, this correlation (i.e., *linking*) results in a shared error component that can exaggerate the size of the difference. While it is not a common practice in education research, there are effect sizes and statistical methods that can account for the dependence of pre-post tests in published data when the correlations are reported.

The bias of **g** is also an issue for researchers who want to measure the impact of interventions on student learning. The efficacy of interventions ranging from curricular designs to classroom technologies have been evaluated and scaled-up based on measures of student learning. For these investigations, it is important to have a measure of student learning that is not excessively dependent on the knowledge that students bring to a class. By using the pooled standard deviation, rather than the maximum possible gain as defined by the pretest, as a standardization coefficient, d avoids the bias toward higher pretest means while accounting for instrument specific difficulty of improving a raw score. We recommend researchers use d rather than **g** for measuring student learning. Besides being the more reliable statistical method for calculating student learning, the use of d by the DBER community would align with the practices of the larger education research community, facilitating more cross-disciplinary conversations and collaborations.

ACKNOWLEDGMENTS

This work was funded in part by NSF-IUSE Grants No. DUE-1525338 and No. DUE-1525115, and is Contribution No. LAA-044 of the International Learning Assistant Alliance.

[1] C. Bereiter, Some persisting dilemmas in the measurement of change, in *Problems in the Measurement of Change*, edited by C. W. Harris (University of Wisconsin Press, Madison, 1963).
[2] L. J. Cronbach and L. Furby, How should we measure "change"—or should we?, Psychol. Bull. **74**, 68 (1970).
[3] R. Hake, Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[4] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, 1977).
[5] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).
[6] S. D. Willoughby and A. Metz, Exploring gender differences with different gain calculations in astronomy and biology, Am. J. Phys. **77**, 651 (2009).
[7] L. Bao, Theoretical comparison of average normalized gain, Am. J. Phys. **74**, 917 (2006).
[8] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).
[9] R. J. Grissom and J. J. Kim, *Effect Sizes for Research: Univariate and Multivariate Applications*, 2nd ed. (Routledge, Abingdon, UK, 2012).
[10] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[11] J. I. Smith and K. Tanner, The problem of revealing how students think: Concept inventories and beyond, CBE Life Sci. Educ. **9**, 1 (2010).
[12] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).
[13] J. V. Korff, B. Archibeque, K. Alison Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50k-student study, Am. J. Phys. **84**, 969 (2016).
[14] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[15] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamela, Toward equity through participation in modeling instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).
[16] I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, Phys. Rev. ST Phys. Educ. Res. **8** (2012).
[17] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[18] S. R. Singer, N. R. Nielsen, and H. A. Schweingruber, *Discipline-Based Education Research* (The National Academies, Washington, DC, 2012).
[19] R. M. Talbot, Taking an item-level approach to measuring change with the Force and Motion Conceptual Evaluation:

An application of item response theory, School Sci. Math. **113**, 356 (2013).

[20] D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas, Front. Psychol. **4**, 863 (2013).

[21] W. P. Dunlap, J. M. Cortina, J. B. Vaslow, and M. J. Burke, Meta-analysis of experiments with matched groups or repeated measures designs, Psychol. Methods **1**, 170 (1996).

[22] S. B. Morris and R. P. DeShon, Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs, Psychol. Methods **7**, 105 (2002).

[23] B. Van Dusen, J.-S. S. White, and E. Roualdes, The Impact of Learning Assistants on Inequities in Physics Student Outcomes, in *Proceedings of the Physics Education Research Conference 2016, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 360–363, arXiv:1607.07121.

[24] J.-S. S. White, B. Van Dusen, and E. A. Roualdes, The Impacts of Learning Assistants on Student Learning of Physics, in *Proceedings of the Physics Education Research Conference 2016, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 384–387, arXiv:1607.07469.

[25] V. Otero, S. Pollock, R. McCray, and N. Finkelstein, Who is responsible for preparing science teachers?, Science **313**, 445 (2006).

[26] https://www.learningassistantalliance.org/.

[27] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[28] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[29] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a smog of bias?, Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[30] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[31] J. Honaker, G. King, and M. Blackwell, Amelia II: A program for missing data, J. Stat. Softw. **45**, 1 (2011).

[32] R. Dou, E. Brewe, J. P. Zwolak, G. Potvin, E. A. Williams, and L. H. Kramer, Beyond performance metrics: Examining a decrease in students? physics self-efficacy through a social networks lens, Phys. Rev. Phys. Educ. Res. **12**, 020124 (2016).

[33] D. B. Rubin, Multiple imputation after 18+ years, J. Am. Stat. Assoc. **91**, 473 (1996).

[34] P. D. Allison, *Missing Data* (Sage, Thousand Oaks, CA, 2002).

[35] E. R. Buhi, P. Goodson, and T. B. Neilands, Out of sight, not out of mind: Strategies for handling missing data, Am. J. Health Behav. **32**, 83 (2008).

[36] C. A. Manly and R. S. Wells, Reporting the use of multiple imputation for missing data in higher education research, Res. High. Educ. **56**, 397 (2015).

[37] Y. Dong and C.-Y. Joanne Peng, Principled missing data methods for researchers, SpringerPlus **2**, 222 (2013).

[38] M. J. Cahill, K. Mairin Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).

[39] T. E. Bodner, What improves with increased missing data imputations?, Structural Equation Modeling: A Multidisciplinary J. **15**, 651 (2008).

[40] J. L. Schafer, Multiple imputation: A primer, Stat. Meth. Med. Res. **8**, 3 (1999).

[41] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *Gender differences in science, technology, engineering, and mathematics (STEM) interest, credits earned, and NAEP performance in the 12th grade* (National Center for Education Statistics, 2015).