Gender gaps and gendered action in a first-year physics laboratory

James Day,[1] Jared B. Stang,[1] N. G. Holmes,[2] Dhaneesh Kumar,[1] and D. A. Bonn[1,*]

[1]*Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z1*

[2]*Department of Physics, Stanford University, Stanford, California, 94305, USA*

(Received 30 January 2015; published 1 August 2016)

[This paper is part of the Focused Collection on Gender in Physics.] It is established that male students outperform female students on almost all commonly used physics concept inventories. However, there is significant variation in the factors that contribute to the gap, as well as the direction in which they influence it. It is presently unknown if such a gender gap exists on the relatively new Concise Data Processing Assessment (CDPA) and, therefore, whether gendered actions in the teaching lab might influence—or be influenced by—the gender gap. To begin to get an estimates of the gap, its predictors, and its correlates, we have measured performance on the CDPA at the pretest and post-test level. We have also made observations of how students in mixed-gender partnerships divide their time in the lab. We find a gender gap on the CDPA that persists from pre- to post-test and that is as big as, if not bigger than, similar reported gaps. We also observe compelling differences in how students divide their time in the lab. In mixed-gender pairs, male students tend to monopolize the computer, female and male students tend to share the equipment equally, and female students tend to spend more time on other activities that are not the equipment or computer, such as writing or speaking to peers. We also find no correlation between computer use, when students are presumably working with their data, and performance on the CDPA post-test. In parallel to our analysis, we scrutinize some of the more commonly used approaches to similar data. We argue in favor of more explicitly checking the assumptions associated with the statistical methods that are used and improved reporting and contextualization of effect sizes. Ultimately, we claim no evidence that female students are less capable of learning than their male peers, and we suggest caution when using gain measures to draw conclusions about differences in science classroom performance across gender.

DOI: 10.1103/PhysRevPhysEducRes.12.020104

I. INTRODUCTION

A consensus has developed in the physics education research literature that a gender gap exists for many of the commonly used physics concept inventories (e.g., FCI, FMCE, BEMA, and CSEM), with male students generally outperforming female students. However, across published studies there is significant variation in the development of the gap over time, the various factors that influence the gap, and even the way in which factors influence the gap (see Ref. [1], which is a review article summarizing 17 different studies). In all likelihood, the observed gender gaps are due to a combination of many gendered factors rather than any one that can be easily modified. For many physics courses—with a variety of learning environments and activities, both in and out of class, and all of the previous life experience of the students—it is difficult to isolate the root causes. This difficulty has contributed to a difference between the actual and desired state of affairs in the physics education research community. There is much that we would like to know that we simply do not yet know. An examination of our ignorance highlights some of the field's needs, and we are in a position to address some of these issues.

The first issue we address is related to a measure of the gender gap. One diagnostic for which the gender gap has not yet been explored is the Concise Data Processing Assessment (CDPA), a relatively new concept inventory that provides a quantitative measure of student abilities related to the nature of measurement and uncertainty and to handling data [2]. We do not know whether a gender gap exists on the CDPA and, if it does, how big the effect is. The second issue we address is related to the analyses of such data, that should tell us about the absence or presence and magnitude of the gender gap. Some commonly used techniques, such as calculating an estimate of gain, are often too simple for a problem as complex as how people learn (which includes but is not limited to acquiring new skills, modifying existing knowledge, and reinforcing specific behaviors) and how learning may differ between genders. We will present a range of alternative analysis methods, including five different metrics for calculating gain. The third issue we address is related to the dynamics of student interactions in the teaching labs and how they might influence, or are influenced by, the gender gap.

*bonn@physics.ubc.ca

Given the right type of data, correlations (or lack thereof) between student performance on a relevant diagnostic and their gendered in-lab behaviors—naturally connected to learning and/or practicing data handling skills—would become apparent. Our interest in a gender gap on the CDPA is closely associated to our concern over how gender plays out in a laboratory environment.

To make a measurement of the gender gap, the CDPA was deployed at the University of British Columbia (UBC), as a pretest and as a post-test, and across several years. To evaluate differences in students' performance on the CDPA, we performed an analysis of covariance (ANCOVA), calculated effect sizes, and contextualized our results. This particular path was chosen after carefully scrutinizing our original approach to the data, one which was similar to many previous studies. To explore how student dynamics might influence student performance on the CDPA, we observed how students spent their time handling the lab equipment, working on the computer, or performing other tasks, and explored relevant correlations.

This paper begins with a brief summary of the theoretical framework that guides our study. We then present the gender gap that exists for the CDPA and work towards an understanding of what our data are telling us. In parallel to a description and application of some standard analyses, we scrutinize some of the issues related to measurements of gain with concept inventories; given the biases introduced by various estimates, we suggest caution when using gain measures to draw conclusions about differences in science classroom performance across gender. We next share the results from an ANCOVA, used to determine the effect of gender on post-test CDPA scores after controlling for pretest CDPA scores. The persistence of the gender gap on the CDPA, from pre- to post-test, raises the question of what might be happening in the lab that could lead to an interaction effect: How do students' gendered dynamics in the lab affect how they ultimately perform on the CDPA post-test? We assess this with observations that compare female and male students' use of lab equipment (presumably to collect data) and computers (presumably to analyze data) in mixed-gender pairs. Finally, we highlight the confounding factors in our work and touch on some potential implications for instructors and future research questions.

II. THEORETICAL BACKGROUND

The theoretical framework we use borrows key ideas from poststructural gender theory [3] and situated cognition [4–6].

By gender, in this paper, we mean the social roles based on the biological sex of the person (culturally learned) and/or the personal identification of one's own gender based on an internal awareness (gender identity). We explicitly acknowledge that the gendering of the discipline of physics is complex in nature; learning physics—and also becoming a physicist—is a gendered experience [7]. Brotman and Moore [8] provide an extensive review of the literature about gender bias in the construction of scientific knowledge, the cultural norms and values of scientific communities, and curricular and pedagogical practices in science education.

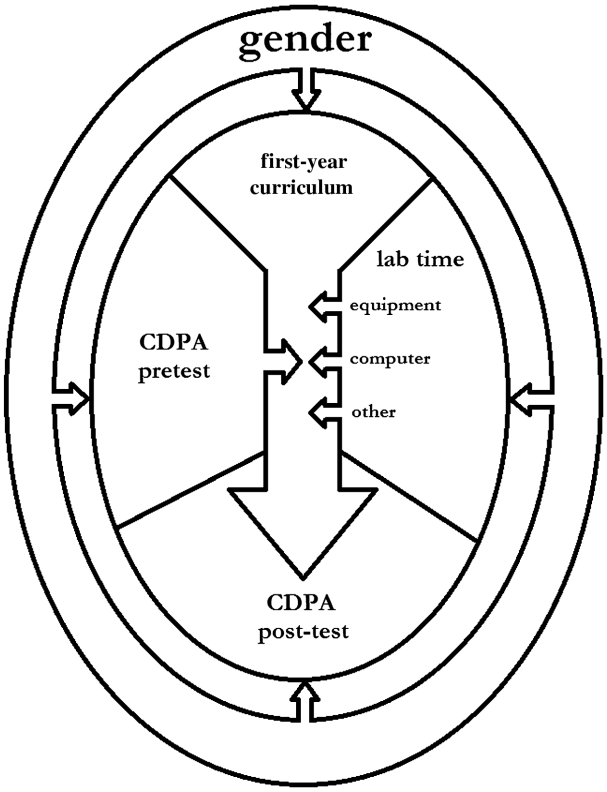

FIG. 1. Schematic representation of the theoretical framework adopted for this study.

In situated cognition [5], knowledge is a product of the activity, context, and culture within which it is developed and used. The enterprise of learning itself is viewed as being contextually constrained; the context limits what can be learned and shapes how it is learned. The concept of a community of practice is fundamental to situated learning [6]. A community of practice is a group of people engaged in a mutual activity, in pursuit of a shared goal. The doing of physics may be understood as participating in a community of practice. The doing of gender may also be understood as participating in a community of (masculine and feminine) practice [9,10]. A second—different but interrelated—perspective of communities of practice pertains to identity, and is concerned with how the individual participants relate to their community of practice. The notion of communities of practice is particularly relevant in our laboratory course by design; we are guiding students to construct their own knowledge and habits of mind in an environment that is nearly 100% collaborative.

We use these systems to make sense of how our physics students perform on a data handling skills diagnostic at the

beginning and end of their first academic year, as well as how they behave in the lab.

The ways in which these theories relate to learning in our lab course are shown in Fig. 1. At a base level, students' learning is guided by the first-year curriculum (upper, inner area), and their interaction with this curriculum should predict how they perform on the post-test (lower, inner area). At some higher level, we recognize that how students assimilate this first-year curriculum depends critically on what they bring with them into their first year, represented by their performance on the pretest (left, inner area). How they assimilate the curriculum is also dependent on how they practice physics, represented by their behavior in the lab (right, inner area). Finally, we acknowledge that ubiquitous and complex gendering always applies (outer boundary). All four of these sections fall under the influence of gender.

III. METHODS

A. Research environment

Our data were collected from the laboratory component of an introductory physics course at the University of British Columbia.

The course is calculus based and is offered as two, consecutive, single-semester labs. It is intended for first-year students with an interest in the science, technology, engineering, and math (STEM) fields. Most of these students will eventually earn a degree in physics, chemistry, or the life sciences. About 80% of the students who take the first semester also take the second semester, while about 5% of the students in the second semester were not registered in the first semester. However, we report only on students who were registered in both semesters, and for whom we have paired pre- and post-test data. Students attend a weekly, three-hour lab section, with 30–50 students in each section, facilitated by two graduate teaching assistants and one instructor. The lecture component associated with this lab covers standard first-year physics material.

The learning goals for this physics laboratory [11] focus on a specific set of foundational experimentation skills. These skills include accounting for the nature of uncertainty in all measurements, developing statistical and graphical methods for evaluating data, and initiating proficiency at collecting and interpreting data. (Similar basic skills can be equally important to those pursuing careers outside of physics, for example, in the medical sciences [12,13].) Such goals are considerably different from those of the traditional first-year physics laboratory, which are often focused on illustrating fundamental concepts and theories, facility with laboratory equipment, and written or oral scientific communication. However, they align well with the AAPT recommendations for the undergraduate physics laboratory curriculum [14], which focuses on constructing knowledge, modeling, designing experiments, developing technical and practical laboratory skills, analyzing and visualizing data, and communicating physics. Physics concepts can be—and are—carefully woven into the course. Substantial research on the extent to which labs can contribute to students' conceptual understanding of physics has been well documented [15]; nevertheless, the primary aim of these laboratories is to establish a meaningful understanding of and a practical mastery of handling data.

Data were collected over five academic years, from 2009–10 to 2013–14. The number of students enrolled ranged from 130 to 145 each year, with a total of 471 having matched pretest and post-test scores. Female student representation in the class ranged from 37% to 44%.

B. Data collection

1. Gender

As mentioned above in Sec. II, we hold the complex gendering of physics to be true. As far as our categorical gender data are concerned, however, we do not—and cannot—properly treat gender as constructed, flexible, and continuous. Instead, we simplify and treat gender as a dichotomous, stable category. We focus on the differences between the genders rather than on the variations within.

Our categorical gender data are generated when an individual first creates a student account as a part of their application to UBC. During the account creation, various biographical data are required: the "First Name" and "Last Name" categories are both open-ended text boxes; the "Date of Birth" categories (day, month, and year) are drop-down menus; the "Gender" category is a radio button selection with only "male" or "female" options available.

2. Concept inventory

Our concept inventory data were obtained using the Concise Data Processing Assessment, a ten-question, multiple-choice instrument that provides a quantitative measure of student abilities related to understanding measurement and uncertainty and to handling data. Specific learning goals targeted by the CDPA include, but are not limited to, being able to weigh the relative importance of numbers that have differing uncertainty; judge whether or not a model fits a data set; linearize exponential distributions, by using semilog plots, and power-law distributions, by using log-log plots and power-law scaling; and extract meaning from the slope and intercept of data that have been linearized. The CDPA was used as both a pretest and a post-test, and was administered during the first (early September) and final (late March) weeks of labs. Students were given 30 minutes to complete all ten questions and were asked not to use a calculator. Their performance was motivated by the explanation that the collective results will help us to improve the quality of the course, as well as the promise that their individual scores could only have an upward influence on

their final lab grade (up to 1% bonus). The CDPA has good evidence of validity, and statistical tests indicate that the CDPA is a reliable assessment tool, with good dynamic range, for measuring targeted abilities in undergraduate physics students. Scores on the CDPA range from about 25% (pretest, for novices) to about 80% (for experts). Details of its development and administration can be found elsewhere [2,16].

We use the CDPA for pedagogical reasons. Beyond having undergone careful item construction, the CDPA probes many of the broadly applicable skills that will be of value to our students regardless of their later academic path. The primary goals of the lab course, described above, also align well to the skills tested by the CDPA. There may be some gender effects specifically related to laboratory instruction and group work in a laboratory environment.

We also use the CDPA for research reasons. Research suggests that more difficult tests produce greater stereotype threat effects [17]. In parallel with our analysis of student performance on the CDPA, we carefully scrutinize the assumptions associated with commonly used statistical methods, the methods themselves, and the contextualization of effect sizes. To best do this, a difficult assessment helps us to achieve the biggest signal.

3. In-lab observations

Our in-lab, student behavior data were collected using a scheme similar to that of two recent studies [18,19] and was inspired by the Baker-Rodrigo Observation Method Protocol [20,21]. On a spatial map of the lab room, observers recorded snapshots of student activity. From an unobtrusive vantage point and while wandering around the room, an observer would record whether a student was handling the lab equipment (coded as *Equipment*); working on their computer (coded as *Computer*); or doing anything else (coded as *Other*), which, in effect, meant writing, talking, being temporarily absent from the room, or off task. Each observation was tagged to a specific seat in the lab room, and each seat corresponded to a specific desktop computer, assigned to a pair of randomly partnered students. This procedure allowed us to obtain a snapshot of the entire classroom (record the instantaneous behavior of each student) in 2–3 minutes. By repeating this process at roughly five-minute intervals, we can construct a timeline of student behaviors over the course of a full lab session.

Two separate observers collected data. One observer (NGH) was a teaching assistant (TA) in two sections of the course while the other observer (JBS) had no previous connection to the course. An observer never served as a TA for any of the lab sections they observed; the lab was always overseen by an instructor and two TAs. In order to establish the reliability of observations, both observers performed a full set of observations for a single lab period. They did not make observations of individuals at the same time; rather, they followed the observation protocol, individually, for the full session. Comparing the codes from the two raters provides a measure of the reliability of the coding scheme for individual students across a full lab session. For each student, a normalized participation in each behavior was calculated from each observer's results. These are the fractions of time that the student was recorded in that behavior when at least one partner in the pair was recorded in that behavior. The Spearman correlation between observers for student equipment use was 0.79, the correlation for student computer use was 0.92, and the correlation for other behaviors was 0.80, all showing good agreement between the two observers.

Observations were collected in all four sections of three separate lab experiments during the second term of the course, of the 2013–2014 academic year. Each female-male pair was observed (a snapshot of their behaviors was recorded) an average of 16 times per lab experiment, and a total of 2133 observations of female-male pairs were made through the three weeks of observations. The three experiments were the first, second, and sixth of ten experiments over the term. The first experiment required making a high-quality measurement of a spring constant using Hooke's law. The second experiment required making a high-quality measurement of the resonant oscillation frequency of the same mass-on-a-spring system. The third experiment required making a high-quality measurement of the time constant of a discharging capacitor in a parallel resistor-capacitor (RC) circuit. These labs offered the best opportunity for observing instances of each type of behavior.

IV. RESULTS AND DISCUSSION

A. Concept inventory

Our assessment data address the open question of whether there exists a gender gap for the CDPA on pretest and on post-test, and whether that gap changes from pre- to post-test.

1. Identifying whether a gap exists

The independent-samples *t*-test is commonly used to determine whether a statistically significant difference exists between the means of two independent groups on a continuous dependent variable. Upon overview of several published research studies in the physics education literature, we have noticed that oftentimes an examination of the underlying assumptions behind a statistical test are not presented (see, e.g., Refs. [22–26], which have collectively been cited more than 250 times). This very well could be a self-selection effect; i.e., only manuscripts containing data fulfilling the assumptions are submitted. But it could also be that violations of assumptions are rarely checked for in the first place—something we were guilty of in the early stages of writing this paper. This is consistent with the findings from a recent study [27] that further revealed a general lack of knowledge about the assumptions themselves, the

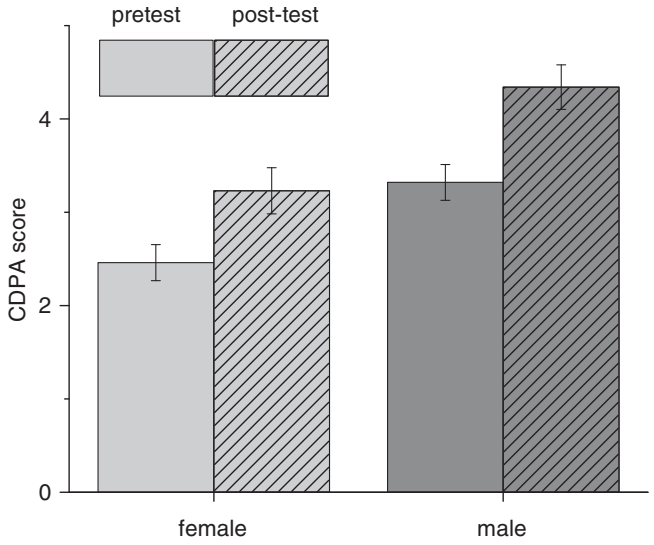

FIG. 2. CDPA pretest (solid columns) and post-test (hatched columns) results for the populations studied. The raw CDPA score is out of a maximum possible of ten points. Our students are learning, but the status quo remains. Uncertainty bars represent the 95% confidence interval. These data consist of 191 female and 280 male students (the ones for which we have paired pre- and post-test data).

robustness of the techniques with regards to the assumptions, and how or whether the assumptions should be checked. Applying any statistical techniques with unmet assumptions can influence both type I and type II errors, as well as result in overestimation or underestimation of inferential measures and effect sizes. Keselman *et al.* [28] argue that "the applied researcher who routinely adopts a traditional procedure without giving thought to its associated assumptions may unwittingly be filling the literature with non-replicable results." To help avoid the fate of nonreplicable results here, the six assumptions that must first be considered in order to run an independent-samples *t*-test are explicitly addressed in Appendix A.

Pre- and post-test scores, for female and male students, are shown in Fig. 2 and Table I. These low scores demonstrate that the CDPA is a difficult assessment; in fact, random guessing produces a score of 23.5%. A summary of the difficulty and discriminatory power of the CDPA is included in Appendix B. On the pretest, there is a statistically significant gender gap favoring males, which has a value of $8.6 \pm 2.7\%$ [mean difference $\pm 95\%$ confidence interval (CI)], $t(469) = 5.95$, $p < 0.001$. On the post-test, there is also a statistically significant gender gap favoring males, which has a value of $(11.1 \pm 2.7)\%$, $t(469) = 6.13$, $p < 0.001$. Examining student post-test performance item by item shows that the gap is fairly uniform across the entire test (see Appendix C). The two exceptions are both questions that require judging the quality of fit of a linear model to data, which are equally difficult for all students.

TABLE I. Summary of CDPA pretest and post-test data shown in Fig. 2.

	CDPA score, with 95% CI		
Gender	Lower bound	Mean	Upper bound
		Pretest	
Female	2.27	2.46	2.65
Male	3.13	3.32	3.51
		Post-test	
Female	2.98	3.23	3.48
Male	4.10	4.34	4.58

That the mean difference is statistically significant, however, is the less interesting thing about the data. This absolute difference does not take into account the variability in scores [29]—after all, not every subject achieved the average outcome. We care not only about whether there is a difference but also about the size of the difference.

An effect size is a quantitative measure of the strength of a particular phenomenon. Knowing the magnitude of an effect allows us to ascertain the practical significance of statistical significance. We use Hedges' g effect size value [30], instead of the more commonly encountered Cohen's d, since the sample sizes of our two groups are unequal. Hedges' g suggests a medium effect size for both our pretest result ($g = 0.56$) and our post-test result ($g = 0.57$). These numbers should be interpreted as group means that differ by slightly more than half a standard deviation, each. In Cohen's terminology [31,32], a small effect size is one in which there is a real effect—something is really happening in the world—but which you can only see through careful study. A large effect size is one that is big enough and/or consistent enough that you may be able to see it "with the naked eye." A medium effect size lies between the above two.

To better contextualize this finding, we can ask what sort of effect sizes other similar studies have found. A recent review of the literature on the gender gap on concept inventories in physics provides a summary of these data: in particular, see Figs. 1 and 2 of Madsen *et al.* [1]. In the seventeen separate studies that they reviewed—of the FCI, FMCE, BEMA, and CSEM—there was (almost) always a gender gap favoring males, on pretests and post-tests. These can be used to calculate effect sizes for each of those seventeen studies, against which we can contrast our effect sizes. Figure 3 shows a histogram of the effect sizes, on pretests and post-tests, for each of the seventeen studies reviewed in Ref. [1] alongside the effect sizes associated with our study. The effect sizes we are observing on the CDPA are at least as large as any found in other similar studies.

Evaluating effect sizes is not easily done. One reason is that most phenomena are multivariable problems: to isolate just one that has an effect on an interesting outcome is a triumph even when, in particular instances, that variable might be overwhelmed by others with opposite influence.

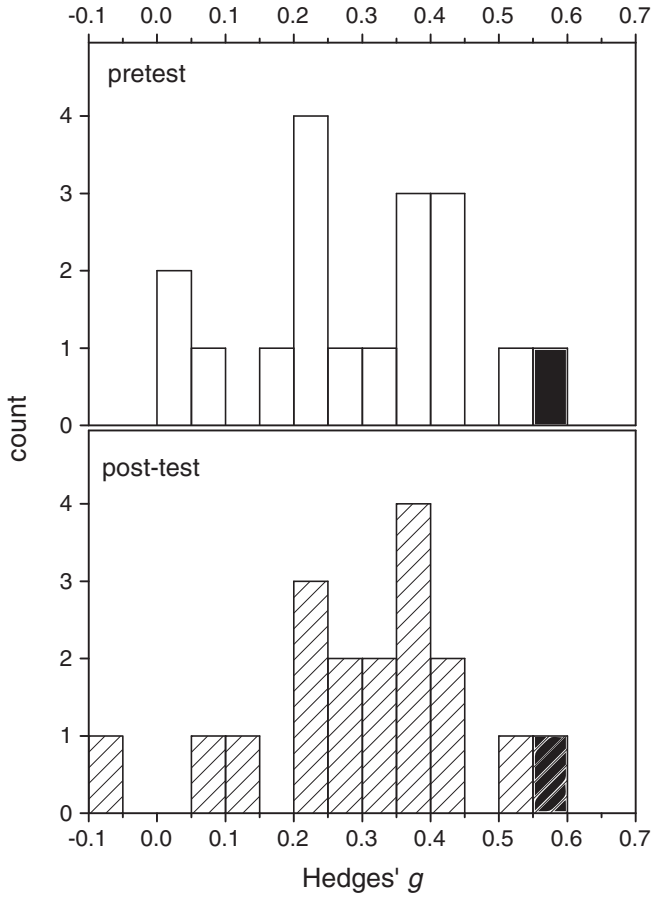

FIG. 3. Histogram of Hedges' g effect sizes for the FCI, FCME, BEMA, and CSEM results (white) (summarized in Ref. [1]) and the CDPA (black). Pretests are given by the solid columns and post-tests by the hatched columns. None of the effect sizes from the studies involving the FCI, FCME, BEMA, and CSEM are as large as the effect sizes associated with this study of the CDPA.

Another reason is that even "small" effects can result in large consequences over time. The CDPA effect sizes indicate that we probably should care about the size of the gap. They are large enough that the difference in performance might be noticed by the students and, therefore, could reinforce negative stereotypes about the women who have already made it this far along the leaky pipeline of STEM [33]. They imply that a sizable minority are performing less well on some rather important skills—this is a theoretical problem, not just a practical one. The perspective of effect size is requisite for informed judgment; without it, we cannot parse that which is firmly relevant from that which is vanishingly subtle and, in some cases, notoriously difficult to replicate.

While our analysis thus far tells us that gaps exist at the pre- and post-test, we further want to explore the impact of our first-year curriculum and students' experiences in the lab on their performance. How does the gap change over time? To explore this, we first examine measures of learning "gain."

2. Gain

Many instructors rely on some measure of gain to quantify how students' performances have changed, from pre- to post-test: a proxy for how much has been learned. But pre- and post-test scores are not exactly measures of the same variable; change, by nature, is multivariate. There is more than one way to measure gain [34], and arguments in favor of using any one of them are, superficially, sensible. But it can be easy to mistake understanding how something is used with whether something should be used. In principle, as researchers, we understand that sufficient context is required when drawing conclusions from one's data. In practice, we have found that this context is significantly less easy to come by when dealing with measures of learning gain.

Recently, the appropriate use of normalized gain and other gain score calculations has been scrutinized by education researchers, noting a strong positive correlation between pretest scores and normalized gain [23]. Given the biases introduced by various learning gains, caution must be taken when using these measures to draw conclusions about differences in science classroom performance across gender. Brogt *et al.* [35] have shown that different expressions for calculating gains have different inherent biases, with Hake's normalized gain [36] being particularly sensitive to high pretest scores.

To emphasize this point, we characterize the gain, from pre- to post-test on the CDPA, using several different estimators. These varied estimates of gain can be biased against different pretest scores.

First, we use the average normalized change [37] $\langle c \rangle$, defined as the ratio of the gain to the maximum possible gain (or as the loss to the maximum possible loss). The equations used to calculate c are given below. Students who score perfectly or zero on both pretest and post-test should be dropped from the data set (none of our students fell into this category). Once normalized changes have been computed for each student, an average normalized change $\langle c \rangle$ can be calculated:

$$c = \begin{cases} (\text{post} - \text{pre})/(\text{max} - \text{pre}) & \text{if post} > \text{pre} \\ 0 & \text{if post} = \text{pre} \\ (\text{post} - \text{pre})/(\text{pre}) & \text{if post} < \text{pre}. \end{cases}$$

Second, we use the average absolute gain $\langle g_{\text{abs}} \rangle$ that has been normalized by the maximum possible test score (to set the dimensions). It has been argued that absolute gain is a more transparent measure to reflect equity across samples within a single population because it does not compensate for different pretest scores, which do exist in cases of inequality [38]. The equation used to calculate g_{abs} is given below; once absolute gains have been computed for each student, an average absolute gain $\langle g_{\text{abs}} \rangle$ can be calculated for each group:

$$g_{\text{abs}} = (\text{post} - \text{pre})/\max.$$

Third, we use Hake's normalized gain [36] $\langle g \rangle$, which is equivalent to the course average normalized gain. This estimator is the most popularly used one in the literature. The importance of Hake's work cannot be overemphasized, as normalized gain has provided many instructors and researchers with a readily accessible and objective measure of performance in their introductory mechanics courses. In this equation, $\langle \text{pre} \rangle$ and $\langle \text{post} \rangle$ are the classes' average pretest and post-test scores out of 100%, respectively:

$$\langle g \rangle = (\langle \text{post} \rangle - \langle \text{pre} \rangle)/(\max - \langle \text{pre} \rangle).$$

Fourth, we use the absolute gain divided by twice the average of the two $\langle g_{2\text{av}} \rangle$, which has been used before [35,39] to demonstrate the potential pitfalls of the estimator of choice. The equation used to calculate $g_{2\text{av}}$ is given below; once absolute gains divided by twice the average of the two have been computed for each student, an average absolute gain divided by twice the average of the two $\langle g_{2\text{av}} \rangle$ can be calculated:

$$g_{2\text{av}} = (\text{post} - \text{pre})/(\text{post} + \text{pre}).$$

Fifth, we use the percent increase over pretest performance $\langle g_{\text{rel}} \rangle$, which is similarly computed from the g_{rel} values of each student. This is the standard definition of relative change:

$$g_{\text{rel}} = (\text{post} - \text{pre})/(\text{pre}).$$

Results from these five metrics of gain are presented in Fig. 4.

We found that male students' average scores resulted in higher apparent learning gains than female students' average scores, but only when $\langle c \rangle$ was used. When $\langle g_{\text{abs}} \rangle$, $\langle g \rangle$, $\langle g_{2\text{av}} \rangle$, and $\langle g_{\text{rel}} \rangle$ are used, the statistical significance goes away and the effect size vanishes (and *maybe* even flips sign). That we find statistical significance in one estimate but not the others suggests to us that none should be trusted here. It is difficult to argue for a verifiable difference in learning gains between female and male students, even if the female students start and end at lower levels of achievement. This type of conflicting pattern has been observed and explained before [39]. These facts highlight that examination of gain scores must be approached with great care and, perhaps, that we are better advised to avoid examining gain scores at all. If different measures of gain are applied to the same raw data and different narratives result, then perhaps, rather than asking questions regarding gain scores, we are better served by framing our question in another way. The question of whether one gender has learned more than another is fraught with unspoken major premises. Instead we can ask whether there is a gendered difference on post-test scores after having controlled for (some of the) differences that female and male students begin the course with.

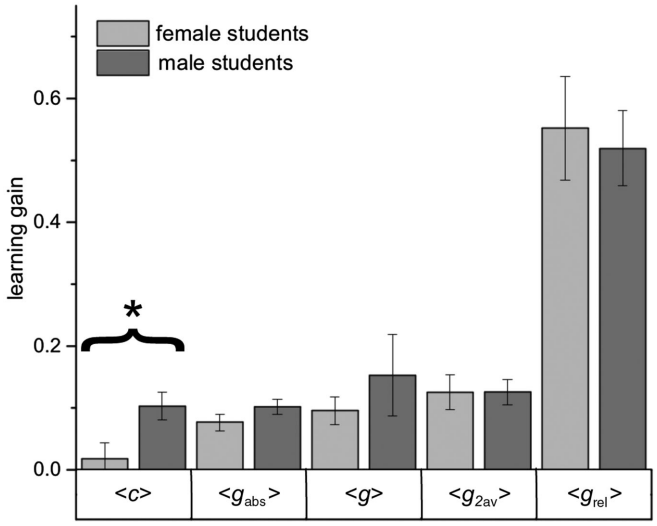

FIG. 4. The "gain" by female and male students on the CDPA, as defined by five different estimators. The asterisk indicates a statistically significant difference below the $p = 0.05$ level from an independent-samples t-test. Uncertainty bars represent the standard errors of the mean. From left to right: (1) the average normalized change $\langle c \rangle$; (2) the average absolute gain normalized by the total test score $\langle g_{\text{abs}} \rangle$; (3) the course average normalized gain $\langle g \rangle$; (4) the absolute gain normalized by twice the average of the pre- and post-test $\langle g_{2\text{av}} \rangle$; and (5) the relative change $\langle g_{\text{rel}} \rangle$. That different estimators lead to different stories suggests that the gain results are suspect.

3. ANCOVA

An analysis of covariance can be used to test the null hypothesis of the equality of two (or more) population means. The assumptions that must first be considered in order to run an ANCOVA are explicitly addressed in Appendix D. An ANCOVA was run to determine the effect of gender on post-test CDPA scores after treating pretest CDPA scores as a covariate. After adjustment for pretest scores, there was a statistically significant difference in post-test scores between genders, $F(1, 468) = 16.86$, $p < 0.001$, partial $\eta^2 = 0.035$, with males scoring higher than females. The adjusted CDPA scores by gender, with pretest as a covariate, are presented in Table II.

TABLE II. Adjusted CDPA scores by gender, with pretest as a covariate; cf. with the unadjusted scores in Table I.

		Adjusted, with 95% CI		
	N	Lower bound	Mean	Upper bound
Female students	191	3.20	3.46	3.72
Male students	280	3.97	4.18	4.40

In summary, the effect of gender is statistically significant but smallish in size, accounting for at least 3.5% of the variance in CDPA post-test scores. The covariate of CDPA pretest score has had a significant impact on the difference we have observed between female and male students, having accounted for as much as 13.1% of the variance in the CDPA post-test scores: how the students finish is largely predicted by how they started.

But small effects can be important in certain contexts. One way in which small effects can be relevant is if they accumulate into larger effects. We know that small effects can add up because students come into the lab with a gender gap that did not exist when they first came into this world. We are measuring over only about 5% of their life so far—that the dominant driver of post-test performance is pretest performance is a testimony to the integral of small effects over time. Small effects can also trigger larger consequences, and we do not want a student awareness of the gender gap on the CDPA—a small effect—to have further-reaching influence and serve as a tipping point for females to disengage in the lab or, worse, depart from physics.

4. Discussion

The gender gap is a complex phenomenon that cannot be easily explained. Many factors very likely contribute to and influence the gender gap, but they are also difficult to observe and quantify. Differential background and preparation are a possible factor. An issue, though, is that it cannot be determined if the differences in the background and preparation variables (which do explain a large portion of the gap observed with other concept inventories) are true differences in preparation or merely artifacts of the testing situation [1].

Stereotype threat is another possible contributing factor. But it is unclear how stereotype threat may affect the gender gap on the CDPA, as there are a number of variables that can moderate stereotype threat effects. Some of these involve the test itself, some involve the test taker, and some involve the testing environment. Research suggests that more difficult tests produce greater stereotype threat effects [17]. The CDPA is an objectively difficult test. Given the crucial role that performance anxiety may play in mediating stereotype threat, it is no surprise that frustratingly hard tests are most likely to induce stereotype threat. Independent of the actual properties of a test, stereotype threat effects can be exacerbated or attenuated by the representation of the test to examinees. Stereotyped groups show greater performance decrements when a test is purported to show intergroup score differences or is represented as diagnostic of ability (see, e.g., Ref. [40]).

Related to our data here, the persistence of the gender gap on the CDPA (i.e., it exists on the pretest and it continues to exist on the post-test) raises the question of what might be happening in the lab that could lead to this interaction effect. To assess this, we used observations that compare female and male students' use of lab equipment (presumably, to collect data) and computers (presumably, to analyze data) in mixed-gender pairs and throughout several lab periods.

B. In-lab observations

Our observation data address the open question of how the dynamics of student interactions might influence the gender gap, or vice versa.

1. Activity participation

To probe the dynamics of student interactions, we studied how behavioral modes differ in mixed-gender pairs. For the academic year in which these observations were collected (2013–2014), student partners were assigned by the course instructor to maximize the number of mixed-gender pairs. These partnerships were constant for the first two weeks of observations, but were different for the third week. Given the gender disparity in enrollment, this meant that there were almost no female-female pairs. For all other years, students have been allowed to partner with whomever they want.

Participation in computer, equipment, and other activities are shown in Table III. Each observation of a pair shows up as one count in the appropriate cell in the table. For example, in 153 of the 2133 distinct observations of mixed-gender pairs, the female partner was using the *equipment* while the male partner was doing some *other* activity. As another example, in 42 of the 2133 distinct observations of mixed-gender pairs, both partners were observed to be using a *computer* at the same time. Collectively, these numbers tell us that female and male students allocate their time differently in the lab. Both genders spend most of their respective time in *other* activities (60.2% for female students and 49.3% for male students). However, while female students roughly split their remaining time between the *computer* (20.6%) and the *equipment* (19.1%), male students tend to spend more of it on the *computer* (29.1%) and less with the *equipment*

TABLE III. Contingency table of gender and in-lab activity, used for the Bhapkar test. Each observation of a pair shows up as one count in the appropriate cell of the table. A similar analysis was done for each week individually, and no noticeable effect exists by topic or as the term goes along. Female and male students do spend their time differently in the lab.

		Female			
		Equipment	Computer	Other	Total
Male	Equipment	137	112	212	461
	Computer	118	42	461	621
	Other	153	286	612	1051
	Total	408	440	1285	2133

(21.6%). Correspondingly, this means that the male student in a female-male pair tends to spend 15 more minutes on the *computer* than his lab partner (52.4 minutes versus 37.1 minutes). Also, this means that the female student in a female-male pair engages in 20 more minutes than her lab partner in *other* activities (108.4 minutes versus 88.7 minutes).

The chi-square test for association is commonly used to determine whether two categorical variables are statistically independent, but it is a test that cannot be used on correlated data. Here, this assumption is violated as we obtain more than one measurement from the same pair. An extension of McNemar's test can instead be used, to test for differences between related variables. While McNemar's test is restricted to two related variables, the Bhapkar test is a powerful extension that allows for larger contingency tables [41,42]. The assumptions relevan to a Bhapkar test are addressed in Appendix E.

The Bhapkar test compares these marginal proportions, and was conducted between gender and participation in *computer*, *equipment*, and *other*. All expected cell frequencies were greater than five. There was a statistically significant association between gender and participation among activities, $\chi^2(2) = 51.7,\ p \leq 0.001$. As discussed above, these differences are primarily in computer use and time spent in other activities.

To help contextualize these numbers, local odds ratios may be calculated. Compared to their male peers, females are 1.73 times as likely to engage in *other* activities versus use a *computer*. Compared to their female peers, males are 1.58 times as likely to use the *computer* versus doing all else combined. Compared to their female peers, males are 1.25 times as likely to use the *computer* versus use the *equipment*. We observe a real but smallish effect in how behavioral modes differ in mixed-gender pairs.

2. Discussion

This analysis makes it clear that female and male students spend their time differently in the lab. It is surprising that no difference existed on equipment use, since a pilot study using this coding scheme did suggest a marginal gendered effect for equipment [19]. We suspect that this would not have been true a generation ago, when males likely would have entered the first-year physics lab with significant experience in "tinkering," whether it be with engines, ham radios, or household gadgets.

Our observations of male students using the computers significantly more than female students does have a confound that we did not address at the time of our data collection: we did not discriminate between personal laptop computers and the in-lab computer towers (one per pair). It is possible that simply more male than female students bring their laptops with them to the lab, and that (naturally) people are reluctant to use someone else's computer. There were lab computers available to all groups, however, so students still had the opportunity to use a computer if they did not bring their own.

C. Relationship between CDPA performance and in-lab actions

Our assessment data revealed a gender gap on the CDPA pretest that persisted on the post-test. Our observation data revealed differences in how female and male students spend their time in the lab. Together, these raise questions about a possible interaction effect. How might students' gendered dynamics in the lab affect how they ultimately perform on the CDPA post-test?

1. Correlations

For each student, a *normalized participation* may be calculated for the *equipment*, *computer*, and *other* behavioral modes. These are the fractions of time that the student was observed in a behavioral mode when at least one partner in their pair was observed in that behavioral mode. A score of zero would mean that the student was never observed using the computer, for example (and the pair was observed to be using the computer at least once). A score of 1 would mean that the student was in sole control of the computer.

With a measure of student performance (CDPA post-test score) and of how much time they spend on the computer (normalized computer time), we can gauge the correlational strength of the relationship between these two variables. However, the normalized computer time is bimodal (lots of high usage and lots of low usage, with little in between), and visual inspection of a normal *Q-Q* plot suggests that we fail to meet the assumption of normality. This fact precludes us from being able to run a Pearson's correlation on our data. Spearman's correlation, though, provides an alternative to obtain a valid and interpretable result.

The first assumption for Spearman's correlation relates to the measurements made: the two variables should be measured at the continuous level. Our variables, the CDPA post-test score and the normalized computer time, were each measured at the continuous level. To establish whether the correlation coefficient is significant, two additional assumptions are required: paired observations and a monotonic relationship between the two variables. Our variables are paired and possess a weak, monotonically increasing relationship.

Spearman's correlation was run to assess the relationship between CDPA post-test score and normalized computer use. We found that there was no correlation between CDPA post-test score and normalized computer use, $r_S(122) = 0.084,\ p = 0.353$.

2. Discussion

We found that there was no correlation between CDPA post-test scores and students' normalized computer use in

the lab. A lack of correlation may be due to a number of variables. First, since introductory students generally obtain low scores on the CDPA (average was less than 50%), we may be witnessing range restriction. Since the range of scores does not span the full spectrum, our correlation is decreased. Alternatively, since the ANCOVA results demonstrated that the size of the gender effect was, in fact, quite small, the gendered relationship between our two variables may be similarly small and, in this case, undetectable. Finally, data handling skills may not be what is learned when students are working on the computer. That is, we did not code for how students spend their time on the computer, so it very well may be that time spent using the computer mostly involves entering data and creating tidy graphical representations. It is reasonable to suppose that most of the practice related to data handling happens when the students are writing in their lab books. Perhaps logbook use is when students were extracting a plausible mathematical model from their data or reconciling data with differing levels of uncertainty. Since time spent writing in a logbook was coded as performing an *other* activity, this information was lost in the coding scheme we used; however, it does leave us with another possible avenue of pursuit. Perhaps there is a gendered difference in how logbooks are used.

Coupled to no difference in how often female and male students handled the equipment in the lab, we take this to mean that in-lab, gendered interactions seem not to influence learning of data handling skills. Of all the things that may affect CDPA post-test performance, we seem to have removed one potential candidate (time on computer) from an innumerably long list. It remains interesting, however, that both of these—apparently unrelated—measures are gendered.

Other correlations might exist. Our students have a rather broad cultural background and it is possible that culturally tempered social characteristics raise barriers to their learning physics in a Canadian lab. Of the four possible combinations, it may matter how Canadians and non-Canadians are partnered with respect to gender. Our students also share a range of native languages and it is possible that performance on the CDPA is consequently hindered. To this end, we are developing a simplified Chinese version of the CDPA (besides English, Chinese is the most common language spoken by our students).

Our present findings force us to modify the theoretical framework presented in Fig. 1. Within this framework, we asserted that gendered, in-lab actions would interact with our first-year curriculum. We do not have the evidence to make this claim. Furthermore, we have underestimated the strength of the influence of the CDPA pretest on the post-test.

Given what we have now measured, our modified theoretical framework is presented in Fig. 5. This schematic is meant to show a few things. First, what the students come into the lab with, as measured by the CDPA pretest, largely determines how they perform on the CDPA post-test. Hypothesizing about what might be responsible for the CPDA pretest results is challenging; they are likely a testimony to the integral of small effects (i.e., gender microaggressions [43]) over time. Second, the first-year lab curriculum contributes to an increase on CDPA scores but not as a function of gender. Third, how students divide their time in the lab is a function of gender but does not seem to impact their performance on the CDPA post-test, although it is possible that an uncoded-for behavior might play some small role.

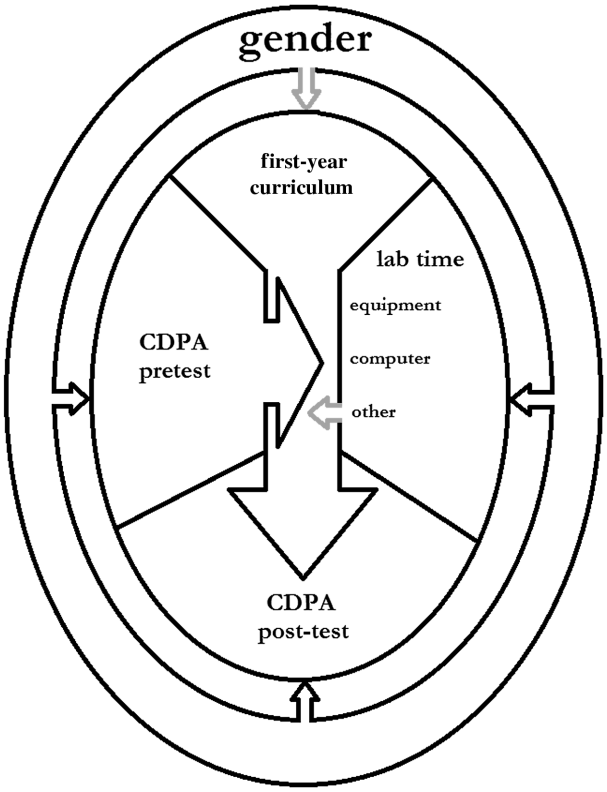

FIG. 5. Schematic representation of the modified theoretical framework, adapted to better align with the results of this study.

V. CONCLUSION

The doing of physics and the doing of gender are inseparable, and must be kept in mind when hoping to make some sense of the gendered experience of learning physics.

The CDPA was used to reveal a statistically and practically significant gender gap at the pre- and post-test level. We can say that everybody learns in our lab, but the status quo (regarding the gender gap) remains at the end of the term. The best predictor of post-test achievement is pretest achievement. Differences in student behavior in the lab were also uncovered: female and male students tend to share the equipment equally, but male students spend more time on the computer and female students spend more time on other activities. However, these gendered actions appear not to correlate with post-test performance measures related to data handling.

In parallel, we have scrutinized some of the commonly employed analysis techniques for the type of data we have collected. The fact that a significant gender effect emerged from the ANCOVA may suggest that the normalized change measure (which was the only other gain measure to provide a significant gender effect) is the most sound. The significant difference in the sizes of the gender effects, however, suggests otherwise. That a method is popular is insufficient reason for its use. We also contend that statistical techniques commonly used in physics education research are often reported with too little information on whether the data being analyzed satisfy the corresponding underlying assumptions: researchers should explicitly check all assumptions so that the use of statistics is never misguided or opportunistic. Further, we argue that researchers use effect size measurements, properly contextualized, to promote clarity in education reform.

We believe that one key reason for the underrepresentation and underperformance of female students in physics [44] is strongly related to their belief in their ability to succeed. Female students generally report lower confidence in themselves than male students. Learning physics requires extensive effortful practice, and having students work in groups can be a valuable tool for effortful practice. The group interactions can provide encouragement and persistence via enjoyable social interactions. To push yourself and get the most out of your practice, one has to believe that success is possible. One way to achieve this belief might be through equivalent peer groups. Equivalent peer groups (for example, matched by GPA and gender) may be a worthwhile endeavor for us, as each group then has approximately the same preparation and motivation, which means one student is not "explaining" to the others. Designed in this way, groups offer a safe environment for the exchange of ideas with peers who are at the same level, thereby allowing for productive conversations to happen. It has been shown that female students profit less than male students from mixed-gender cooperative learning in physics, especially where problem solving is involved [45]. It has also been shown that groups where the members have equal ability result in the most productive interactions and learning [46]. Since our data show that female students are participating in the lab differently than male students in mixed-gender groups, equivalent peer groups may help balance participation—after all, participation in any lab activity is a zero-sum game. This may subsequently affect female student confidence, leading to changes in the CDPA scores.

Other open questions that come from this work are related to when practice with data handling actually happens. Are students actually practicing their data handling skills? If so, is the logbook where students practice them? How is time on the computer spent? These questions should be investigated in future studies. Finally, the meaning of the gender gap on the CDPA and the nature of gendered actions in the lab should not be categorized as well understood.

ACKNOWLEDGMENTS

The authors thank Catherine Rawn for her guidance with our statistical tests. This work has been supported by the Department of Physics and Astronomy at the University of British Columbia, through the Carl Wieman Science Education Initiative.

APPENDIX A: ASSUMPTIONS FOR THE INDEPENDENT-SAMPLES *t*-TEST

The first three assumptions relate to the study design and the measurements made. One must have a dependent variable that is measured at the continuous level; an independent variable that consists of two categorical, independent groups; and independence of observations, which means that there is no relationship between the observations in each group of the independent variable or between the groups themselves. All three of these assumptions are met. Our dependent variable, measured at the continuous level, is the raw score obtained on the CDPA (at pre- or post-test). Our independent variable, of two categorical and independent groups, is the self-identified gender of the students. And our observations within each sample are independent (they do not influence each other—collecting a CDPA score for a male student does not influence the CDPA score for any other male student or for any female student).

The next three assumptions relate to the characteristics of the data collected. One must have no significant outliers in the two groups of your independent variable in terms of the dependent variable; an approximately normally distributed data set for each group of the independent variable; and homogeneity of variances, which means that the variance is equal in each group of the independent variable. Our CDPA data sets contain no significant outliers (i.e., equal to or below $Q1 - 3 \times \mathrm{IQR}$, or equal to or above $Q3 + 3 \times \mathrm{IQR}$, where IQR is the interquartile range). Our CDPA data sets are approximately normal, as assessed by visual inspection of normal Q-Q plots—independent-samples t-tests are anyway robust to violations of normality. There was homogeneity of variance on the pretest, as assessed by the variance ratio [47], Hartley's $F_{\max} = 1.46$. There was also homogeneity of variance on the (paired) post-test, as assessed by the variance ratio [47], Hartley's $F_{\max} = 1.39$.

APPENDIX B: CLASSICAL TEST METRICS FOR THE CDPA

The difficulty, discriminatory power, and reliability of the CDPA have been presented before [2]; those results—which focus both on item analysis (item difficulty index, item discrimination index, and point-biserial coefficient) and on the entire test (test reliability and Ferguson's delta)—are summarized in Table IV. Scores on the CDPA range from chance (for novices) to about 80% (for experts), indicating that it possesses good dynamic range. These results indicate

TABLE IV. Summary of statistical post-test results for the CDPA.

Test statistic	Reasonable Lower bound	CDPA Value
Item difficulty index, P	≥ 0.3	0.39
Item discrimination index, D	≥ 0.3	0.43
Point-biserial coefficient, r_{pb}	≥ 0.2	0.21
Ferguson's delta, δ	≥ 0.9	0.94
Test-retest stability (Pearson)	≥ 0.7	0.80

that the CDPA is sufficiently reliable for the purposes of probing how well students actually handle data.

APPENDIX C: BY ITEM AND BY GENDER PERFORMANCE

Student performance, by item and by gender, is shown in Fig. 6. Questions 5 and 6, which require judging the quality of fit of a linear model to data, are equally difficult for all students. Otherwise, the gap is fairly uniform across the remaining items.

APPENDIX D: ASSUMPTIONS FOR THE ANCOVA

While an analysis of variance (ANOVA) can also be used for this purpose, an ANCOVA carries a few advantages. The first advantage is that it has better ability in finding a significant difference between groups—when one exists—by reducing the within-group error variance. The second advantage is that it reduces bias associated with a chance difference (or differences) previously existing between groups. The ANCOVA captures these advantages by inclusion of a covariate measurement (or measurements). These covariates are incorporated into the ANCOVA analysis such that the ANCOVA error term is almost always smaller (often by a fair bit) than the corresponding ANOVA error term. Furthermore, the dependent variable means are adjusted to partially account for previously existing chance differences between the groups.

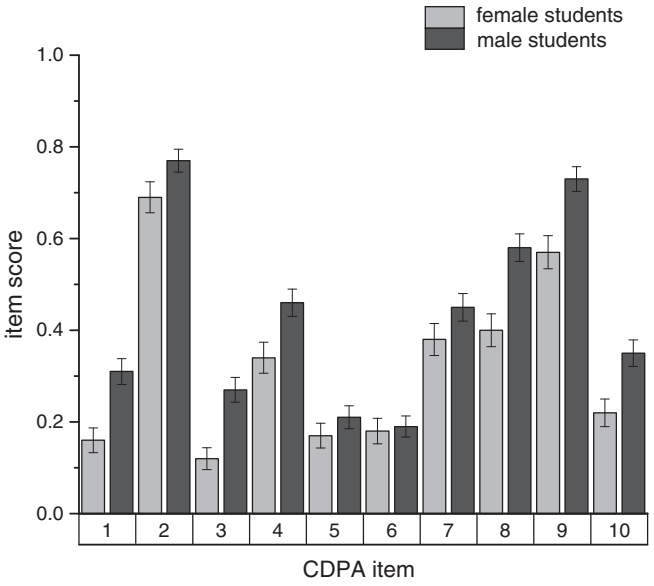

FIG. 6. Summary of post-test performance on each item of the CDPA, by gender. Uncertainty bars represent the standard errors of the mean. These data consist of 191 female and 280 male students (the ones for which we have paired pre- and post-test data). Item by item, the male student population outperforms the female student population.

An ANCOVA requires that multiple assumptions be met for valid and interpretable results. The first three assumptions are shared with all linear models: the residuals are normally distributed; there are no outliers in the data; and there is homogeneity of variance (i.e., the variance of the residuals is equal for the different groups of the independent variable). Three important additional considerations are there is homogeneity of regression slopes; the covariate is linearly related to the dependent variable at each level of the independent variable; and there is homoscedacity (i.e., the variance of the residuals is equal for all predicted values). A final, important assumption is that the covariate and the independent variable are independent from one another. The problem of the covariate and independent variable sharing variance is common, and is misunderstood by many [48]. Our covariate of CDPA pretest score certainly shares variance with CDPA post-test score, as both depend on gender. That our groups differ on the covariate means that our partial η^2 result should be thought of as a lower bound rather than absolutely.

There was normality of standardized residuals of CDPA post-test scores for each gender, as assessed by visual inspection of normal Q-Q plots. (An ANCOVA is anyway fairly robust to deviations from normality. The central limit theorem means that as sample sizes get larger, the less the assumption of normality matters because the sampling distribution will be normal regardless of what our population or sample data look like [49].) There were no significant outliers in the data (i.e., equal to or below $Q1 - 3 \times \text{IQR}$, or equal to or above $Q3 + 3 \times \text{IQR}$). There was homogeneity of variance, as assessed by the variance ratio [47], Hartley's $F_{\max} = 1.39$. There was homogeneity of regression slopes; as the interaction term was not statistically significant, $F(1, 467) = 0.382$, $p = 0.540$. There was a linear relationship between pre- and post-test of the CDPA for each gender, as assessed by visual inspection of a scatter plot.

A linear regression established that pretest scores could statistically significantly predict post-test scores: for male students, $F(1, 278) = 47.97$, $p < 0.001$, and the pretest score accounted for 14.4% of the explained variability in the post-test score; for female students, $F(1, 189) = 21.23$, $p < 0.001$, and the pretest score accounted for 9.6% of the explained variability in the post-test score. The regression equations were as follows: for males, CDPA post-test score $= 2.75 + 0.48 \times (\text{CDPA pretest score})$; for female

students, CDPA post-test score $= 2.23 + 0.41 \times$ (CDPA pretest score). There was homoscedasticity, as assessed by visual inspection of a scatter plot.

APPENDIX E: ASSUMPTIONS FOR THE BHAPKAR TEST

As it is an extension, the Bhapkar test must satisfy the same assumptions required of the McNemar test [50]. The first two assumptions relate to the characteristics of the data themselves: one must have a categorical dependent variable with more than two categories and a categorical independent variable with two (or more) related groups, and the groups of the dependent variable are mutually exclusive. The third assumption relates to how the data were collected: the cases are a random sample from the population of interest.

All three of these assumptions are met. Our dependent variable, measured at the categorical level, is the observed behavioral mode of the student (*equipment*, *computer*, and *other*). Our independent variable, of two categorical and mutually exclusive groups, is the self-identified gender of the students. And our observations were collected randomly (they occurred at roughly five-minute intervals but were completely independent of the timing and content of the lab instruction).

[1] 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).

[2] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).

[3] J. Butler, *Gender Trouble* (Routledge, New York, 1999).

[4] J. S. Brown, A. Collins, and P. Duguid, Situated cognition and the culture of learning, Educ. Res. **18**, 32 (1989).

[5] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation* (Cambridge University Press, Cambridge, England, 1991).

[6] M. K. Smith, Jean Lave, Etienne Wenger and communities of practice, www.infed.org/biblio/communities_of_practice.htm.

[7] A. T. Danielsson and C. Linder, Learning in physics by doing laboratory work: towards a new conceptual framework, Gender Educ. **21**, 129 (2009).

[8] J. S. Brotman and F. M. Moore, Girls and science: A review of four themes in the science education literature, J. Res. Sci. Teach. **45**, 971 (2008).

[9] C. Paechter, *Women's Studies International Forum* (Elsevier, New York, 2003), Vol. 26, pp. 69–77.

[10] C. Paechter, *Being Boys, Being Girls: Learning Masculinities and Femininities* (McGraw-Hill Education, Maidenhead, Berkshire, 2007).

[11] http://www.phas.ubc.ca/~phys109/LearningGoals.html.

[12] S. L. Sheridan and M. Pignone, Numeracy and the medical student's ability to interpret data, Effect. Clin. Pract.: ECP **5**, 35 (2002).

[13] L. M. Schwartz, S. Woloshin, W. C. Black, and H. G. Welch, The role of numeracy in understanding the benefit of screening mammography, Ann. Intern. Med. **127**, 966 (1997).

[14] http://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.

[15] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

[16] https://www.physport.org/assessments/Assessment.cfm?I=55&A=CDPA.

[17] C. M. Steele, S. J. Spencer, and J. Aronson, Contending with group image: The psychology of stereotype and social identity threat, Adv. Exp. Soc. Psychol. **34**, 379 (2002).

[18] J. B. Stang and I. Roll, Interactions between teaching assistants and students boost engagement in physics labs, Phys. Rev. ST Phys. Educ. Res. **10**, 020117 (2014).

[19] N. G. Holmes, I. Roll, and D. A. Bonn, Participating in the physics lab: Does gender matter?, Phys. Can. **70**, 84 (2014).

[20] J. Ocumpaugh, Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual Version 1.0 (Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences, 2012).

[21] J. Ocumpaugh, R. Baker, S. Gaudino, M. Labrum, and T. Dezendorf, in *Artificial Intelligence in Education*, Lecture Notes in Computer Science Vol. 7926, edited by H. Lane, K. Yacef, J. Mostow, and P. Pavlik (Springer, Berlin, 2013), pp. 624–627.

[22] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2002).

[23] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[24] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[25] 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).

[26] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, G. L. Cohen, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).

[27] R. Hoekstra, H. A. L. Kiers, and A. Johnson, Are assumptions of well-known statistical techniques checked, and why (not)?, Front. Psychol. **3**, 137 (2012).

[28] H. J. Keselman, C. J. Huberty, L. M. Lix, S. Olejnik, R. A. Cribbie, B. Donahue, R. K. Kowalchuk, L. L. Lowman, M. D. Petoskey, J. C. Keselman, and J. R. Levin, Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses, Rev. Educ. Res. **68**, 350 (1998).

[29] G. M. Sullivan and R. Feinn, Using Effect Size—or Why the *P* Value Is Not Enough, J. Grad. Med. Educ. **4**, 279 (2012).

[30] L. V. Hedges, Distribution theory for glass's estimator of effect size and related estimators, J. Educ. Stat. **6**, 107 (1981).

[31] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).

[32] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge Academic, New York, 2013).

[33] J. C. Blickenstaff, Women and science careers: Leaky pipeline or gender filter?, Gender Educ. **17**, 369 (2005).

[34] L. Törnqvist, P. Vartia, and Y. Vartia, How should relative changes be measured?, Am. Stat. **39**, 43 (1985).

[35] E. Brogt, D. Sabers, E. E. Prather, G. L. Deming, B. Hufnagel, and T. F. Slater, Analysis of the astronomy diagnostic test, Astron. Educ. Rev. **6**, 25 (2007).

[36] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[37] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).

[38] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. OBrien, I. Rodriguez, and P. Pamelá, Toward equity through participation in Modeling Instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).

[39] S. D. Willoughby and A. Metz, Exploring gender differences with different gain calculations in astronomy and biology, Am. J. Phys. **77**, 651 (2009).

[40] C. M. Steele and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Pers. Soc. Psychol. **69**, 797 (1995).

[41] A. Agresti, *Categorical Data Analysis* (Wiley, Hoboken, New Jersey, 2013).

[42] V. P. Bhapkar, A note on the equivalence of two test criteria for hypotheses in categorical data, J. Am. Stat. Assoc. **61**, 228 (1966).

[43] https://www.unh.edu/sites/www.unh.edu/files/departments/unh_advance/PDFs/microaggressions.pdf.

[44] L. E. Kost-Smith, University of Colorado, Ph.D. thesis, 2011.

[45] E. Harskamp, N. Ding, and C. Suhre, Group composition and its effect on female and male problem-solving in science education, Educational Research **50**, 307 (2008).

[46] Y. Lou, P. C. Abrami, J. C. Spence, C. Poulsen, B. Chambers, and S. d'Apollonia, Within-class grouping: A meta-analysis, Rev. Educ. Res. **66**, 423 (1996).

[47] E. S. Pearson and H. Hartley, *Biometrika Tables for Statisticians* (Cambridge University Press, Cambridge, 1954).

[48] A. Miller and J. P. Chapman, Misunderstanding analysis of covariance, Journal of abnormal psychology, **110**, 40 (2001).

[49] A. Field, *Discovering Statistics Using SPSS* (Sage Publications, Los Angeles, 2009).

[50] For membership holders: https://statistics.laerd.com/premium/spss/mt/mcnemars-test-in-spss-3.php, for non-membership holders https://statistics.laerd.com/spss-tutorials/mcnemars-test-using-spss-statistics.php.