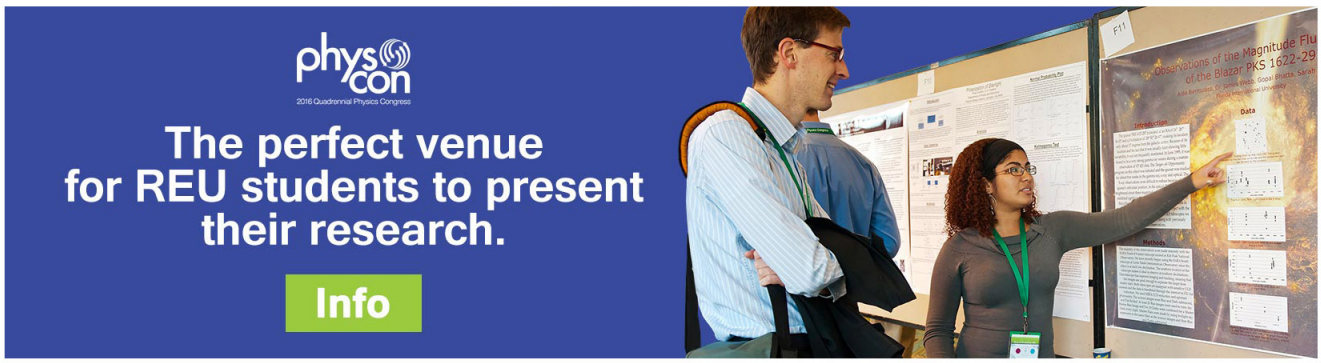
AAPT PHYSICS EDUCATION | AMERICAN JOURNAL of PHYSICS

Exploring gender differences with different gain calculations in astronomy and biology

Shannon D. Willoughby and Anneke Metz

Citation: American Journal of Physics **77**, 651 (2009); doi: 10.1119/1.3133087
View online: http://dx.doi.org/10.1119/1.3133087
View Table of Contents: http://scitation.aip.org/content/aapt/journal/ajp/77/7?ver=pdfcov
Published by the American Association of Physics Teachers



PHYSICS EDUCATION RESEARCH SECTION

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

Exploring gender differences with different gain calculations in astronomy and biology

Shannon D. Willoughby[a)]
Department of Physics, Montana State University, Bozeman, Montana 59717

Anneke Metz
Department of Cell Biology and Neuroscience, Montana State University, Bozeman, Montana 59717

(Received 22 January 2008; accepted 21 April 2009)

To investigate differences in learning gains by gender, we collected data in large introductory astronomy and biology courses. Male astronomy students had significantly higher pre- and post-test scores than female students on the astronomy diagnostic test. Male students also had significantly higher pretest and somewhat higher post-test scores than female students on a survey instrument designed for an introductory biology course. For both courses, 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. Implications for using different learning gain measures in science classrooms, as well as for research on learning differences by gender are discussed. © *2009 American Association of Physics Teachers.*
[DOI: 10.1119/1.3133087]

I. INTRODUCTION

There has been much work in physics education research (PER) on instructional practice.[1] Survey instruments used to assess student mastery of course material have been developed, including the Force Concept Inventory (FCI) (Ref. 2) and Astronomy Diagnostic Test (ADT).[3] Both the FCI and ADT are nationally utilized, reliable, and validated and have been used to gauge the effectiveness of various teaching methods in their disciplines.[3–7]

Education research in biology is not as mature, and appropriate instruments for evaluating learning in the biological sciences are still being developed.[8] A diagnostic instrument modeled after the FCI, the Biology Concept Inventory (BCI) is currently undergoing validation studies.[9,10] Several other instruments, for example, the Student Assessment of Learning Gains,[11] are designed to assess biology curricula and to examine student perceptions of learning. Bowers *et al.*[12] designed a knowledge survey (KS) for an introductory biology course, with questions on material typically covered in lecture and to varying levels of Bloom's Taxonomy of Learning Objectives.[13] The KS queried student's perceived confidence level with the material rather than testing student knowledge itself. Commercially available subject exams, for example, the Major Field Test, administered by the Educational Testing Service,[14] are available but are inappropriate for gauging learning in a single course.

Gender gaps or differences in average performance between men and women have been demonstrated with several standardized tests, particularly those involving science and mathematics.[15,16] Instruments that have been used in the assessment of university science courses (and where women consistently underperform) include the Force and Motion Concept Evaluation (FMCE),[17] Brief Electricity and Magnetism Assessment (BEMA),[18] FCI,[19,20] and the ADT.[3,6] This performance gap holds whether these instruments are used as pre- or post-tests.

Although less educational research has been done using biology learning instruments, there is evidence that women also underperform men on standardized biology tests. For instance, women's scaled scores on the biological sciences component of the Medical College Admissions Test (MCAT) continually lag behind men's scores (8.9 for males and 8.2 for females in 2005).[21] Comparison of MCAT scores back to 1991 indicates that this gender gap has remained unchanged for at least 15 years.[22]

Several reasons have been put forth to explain the effect of gender on standardized test performance, including the participants' perception of their gender roles and the perception of tasks as stereotypically masculine or feminine.[23] Performance gaps may also be due to gendered cues from the questions themselves.[20,23] McCullough's work indicates that women's inferior FCI performance may be due in part to a masculine context bias.[20] We return to McCullough's work in Sec. IV.

Much of the assessment of learning in the classroom utilizes learning gain scores calculated from pre- and post-test administrations of a learning instrument such as the FCI[4,5,18,19,24] and ADT.[6,25] The most commonly used method for calculating gains is the normalized gain $\langle g \rangle$ that measures the fraction of the maximum possible gain: (post − pre)/(100% − pre).[4] This expression implies that higher pretest scores will result in disproportionately higher gain

values than lower pretest scores if the absolute gains are equal. The effect of this skewness has been noted. Brogt *et al.*[25] found that the normalized gain disproportionally rises with higher pretest scores and found larger $\langle g \rangle$ values associated with higher pretest scores, both in their local ADT data sets and in dummy data sets generated by three models of learning gains.

Calculated $\langle g \rangle$ values have been widely used to quantify classroom learning.[4,18,19,26] Given the demonstrated and consistent gender performance gap on many standardized learning instruments and the tendency of the normalized gain measure to favor higher pretest scores, it is expected that calculated $\langle g \rangle$ values would be lower for women than for men. Lower $\langle g \rangle$ values have been reported for female compared to male students on the FCI.[19,27] Lorenzo *et al.*[19] reported a lower $\langle g \rangle$ value for women in a partially interactive physics course ($\approx$0.52 for women and $\approx$0.58 for men)[19] (their Fig. 2) even when the absolute gain on the FCI was higher for women in that course (20.6 for women and 15 for men; data in their Table II and from Ref. 28). This observation does not invalidate the work of Lorenzo *et al.*,[19] who claimed only that interactive teaching methods reduce the gender gap between pre- and post-test administrations of the FCI. This example illustrates that on tests that have demonstrated persistent gender gaps, the value of $\langle g \rangle$ might make it appear that men have learned more than women even when women, on average, post larger absolute gains in the same course.

We are interested in exploring whether pre- and post-test data and gain calculations are useful in indicating whether female students learn less or more in science courses than their male counterparts, independent of their level of incoming knowledge about the subject matter of the course. Because males often begin science courses with more incoming knowledge on the subject (as indicated by higher pretest scores), is it reasonable to conclude that they have learned more in the course if their normalized gain is higher than that of females? Or is the gap in normalized gain scores an artifact of the gender difference seen in pretest scores?

To answer these questions we examined learning gains in both astronomy and biology. The astronomy course study utilized the ADT. Because no appropriate nationally validated instrument was available,[29] biology course instructors developed a diagnostic test [biology diagnostic test (BDT)] to measure learning in the introductory biology course. Given that Brogt *et al.*[25] argued that identical pre- and post-test data sets can give rise to conflicting learning gain results depending on the gain calculation model used, we were particularly interested in examining how the methods compared in terms of showing any differences in learning gains by gender and how learning gain data can be used to make conclusions about gender differences in learning.

II. METHODS

A. Instrument design

The Astronomy Diagnostic Test v. 2.0 (Ref. 3) was administered as a pre- and post-test to students enrolled in Physics 101 [the nonmajor introductory astronomy course at Montana State University (MSU)] during the Spring and Fall semesters of 2006, 2007, and the Fall semester of 2008 (five semesters in total). This test covers topics typically taught in middle and high school science courses and includes nightly and yearly apparent stellar motion, gravitational forces, and phases of the moon.

A 34-item BDT was developed by the course instructors (C. Palen, C. Morrison, and Metz) to measure student learning in Biology 214 (Introduction to cell biology and genetics) and was administered in Fall 2006 and 2007. The instrument was designed to gauge the level of cell biology knowledge at the start of the course and is consisted of 12 factual and 22 conceptual questions about diffusion and osmosis, cell structure, metabolism, and genetics—concepts we expected strong students to have learned in a good high school level biology course. Questions covered the first four levels of Bloom's Taxonomy[13] (knowledge, understanding, application, and analysis) and simultaneously served as a test of student learning of the basic concepts taught in the course.

Because the BDT is not a nationally validated instrument, preliminary validation was performed by three introductory biology instructors at MSU before being field tested by 213 biological science majors over two semesters. We gauged the difficulty of a question by the percentage of correct responses on the pretest, which averaged 47% (standard deviation of 21.5%). This average is higher than published pretest averages for the ADT ($\approx$32%) (Ref. 3) and the FCI ($\approx$39%),[4] but validated instruments can have a wide range of average pretest scores. Kaplan and Saccuzzo[30] suggested that multiple choice questions ideally have initial correct response averages about halfway between random guess and 100%, which for the BDT would have been about 60%. We also tested the reliability of the BDT by calculating the Pearson correlation coefficients between two separate administrations of the BDT as a pretest and a post-test (in Fall 2006 and 2007). The correlation between pretest scores of the first and second administration was 0.97 and between post-test scores was 0.99. Thus this instrument appears to have a very high test-retest reliability. As another test of validity we examined the correlation between post-test scores and course grade. The Pearson correlation coefficient is 0.70, a reasonable figure given that only about 25% of students that took the post-test received a grade of C+ or lower in the course.

B. Data collection

Students in Physics 101 took the ADT on the first day of class and again during the last week of class. Combined enrollment of all sections was close to 2000, but only 1133 (57% overall, 633 male and 500 females) students were present for both testing dates (only students present for both test administrations are included in the study). Pre- and post-test scores were paired for each student to allow for a more robust statistical analysis and allowed us to calculate the gain score for each student (as suggested in Ref. 31).

For the biology course the pre-test was administered as a homework assignment on WEB-CT course management software. Students were required to access the BDT in the first 2 days of the semester and had 40 min to complete the survey. Students were given participation points for completing the instrument, which was not graded and not returned. For the post-test the instrument was embedded in the final exam. Of 231 students completing the biology course over two semesters, 213 (92%) paired pre- and post-test BDT scores (105 males and 108 females) were collected.[32]

C. Data analysis

All statistical data analysis was performed with MINITAB, Version 15. A two-sample t-test was performed on the pre- and post-test data to determine if males had a statistically significant difference in mean scores on their pretest score and to determine the variance of each set of data. Prior to analysis, each data set was examined for normality by skewness and kurtosis calculations.

Regression analysis was done on the data to determine the relation between pre- and post-test scores. Both regression lines follow the relation, post-test$=a_0\times$pretest$+a_1$, with a_0 and a_1 being nonzero. The astronomy and biology models were compared to learning models proposed in Ref. 25 to determine the most likely calculation biases. Average course gains were calculated using several different methods[33–35] in addition to the absolute gain G_0,

$$G_0 = (\text{post-test} - \text{pretest}), \tag{1}$$

$$G_1 = (\text{post-test} - \text{pretest})/(100\% - \text{pretest}), \tag{2}$$

$$G_2 = (\text{post-test} - \text{pretest})/(\text{post-test} + \text{pretest}), \tag{3}$$

$$G_3 = (\text{post-test} - \text{pretest})/\text{pretest}, \tag{4}$$

where G_1 is equivalent to the course average normalized gain $\langle g_{\text{av}}\rangle$,[4] defined as the average of the single student normalized gains; G_1 differs slightly from the course average normalized gain $\langle g\rangle$, which is the average actual gain divided by the average maximum possible gain after averaging pre- and post-test scores for the class as a whole. Because pre- and post-test scores were paired, we calculated $\langle g_{\text{av}}\rangle$. For classes with more than 20 students, these two types of averages are generally within 5% of one another.[4] The values of $\langle g\rangle$ and $\langle g_{\text{av}}\rangle$ for our data differ by 2.9% for the ADT and 0.6% for the BDT and did not affect the outcome of our comparisons.

G_2 is the difference in the pre- and post-tests divided by twice the average and is the only gain calculation in this study that is symmetric about the mean. G_3 has also been used in literature and reflects the percent increase over pretest performance.[7]

Covariate analysis (ANCOVA), used to determine if each gain type depended on gender, pretest score, and a combination of the two, was tested with a model that consisted of linear terms in the pretest score and gender and an interaction term which was the product of the two.

Student grade performance was determined for each course and grades by men and women were compared. Average grades for men and women were determined using data from both semesters of the biology course ($N=213$) and two semesters of the astronomy course ($N=683$) and compared using two-sample t-tests.

III. RESULTS

Students showed significant increases in diagnostic test scores in both biology and astronomy (see Table I). Student scores nearly doubled on the BTD (79% increase, $p<0.0005$, paired t-test) and increased by 44.2% in astronomy ($p<0.0005$, paired t-test). Student performance on the ADT was consistent with national averages; in a national study the average pretest and post-test ADT scores were 32.4% and 47.3%, respectively.[3]

Table I. Overall diagnostic test scores (in percentages) in biology and astronomy before and after completing the course. The standard error of the mean is shown in parentheses.

Course	Mean pretest	Mean post-test	Paired t-test p-value
Biology ($N=213$)	46.9(0.96)	83.9(0.82)	<0.0005
Astronomy ($N=1133$)	34.3(0.43)	49.7(0.54)	<0.0005

Male students outperformed female students on both diagnostic instruments on the pre- and post-tests (Table II). The differences in astronomy were statistically significant, with $p<0.0005$ in each instance. Both nationally and at MSU men outperform women by approximately 10 percentage points on the ADT given as a pre- or post-test.[3] The performance gap between male and female biology students was more modest, with men outscoring women by 3–4 percentage points on both pre- and post-test administrations of the BDT ($p=0.03$ for pretest and $p=0.07$ for post-test). These differences did not arise from differences in overall class performance in either course. (Grades in both courses are determined by a combination of exam scores, homework/laboratory assignments and participation in in-class activities.) In astronomy the final course average was 75.9% for males and 76.2% for females ($p=0.794$); in biology the averages were 76.6% for males and 76.4% for females ($p=0.93$). Males and females also showed no differences in individual exams (including the final exam) and in-class assignments (t-test, $p<0.0005$ in all cases).

Despite the significant gender gaps in pre- and post-test administrations of these instruments, the differences in absolute gains between men and women in both disciplines were modest or nonexistent. The average absolute gain between pre- and post-test scores on the ADT was 16.3 percentage points for men and 14.4 percentage points for women. This difference of less than 2 percentage points between the two groups is statistically significant (two-sample t-test, $p=0.017$), but is only a fraction of a question (0.4) on the ADT. There was no demonstrable difference in the average absolute gain on the BDT across gender (36.4 percentage point change for males and 37.6 percentage point change for females, with $p=0.982$).

Learning gains are often calculated by more complex relations than the absolute gain, including G_1 (normalized

Table II. Mean performance of male and female students on the ADT and BDT, plus or minus the standard deviation. (Standard deviations were not published for the ADT national data.) Differences between males and females were significant (t-test): BDT pretest, $p=0.030$; BDT post-test, $p=0.070$; ADT pretest, $p<0.0005$; and ADT post-test, $p<0.0005$.

Gender (diagnostic)	N	Mean pretest (%)	Mean post-test (%)	Mean G_0
Female, MSU (ADT)	500	28.2 ± 11.8	42.6 ± 16.5	14.4
Female, National (ADT)[a]	≈2700	27	41.5	14.5
Male, MSU (ADT)	633	39.0 ± 14.6	55.3 ± 17.5	16.3
Male, National (ADT)[a]	≈2600	38	53.7	15.7
Female (Biology)	108	44.9 ± 14.4	82.5 ± 12.5	37.6
Male (Biology)	105	49.0 ± 13.4	85.4 ± 11.3	36.4

[a]Reference 3.

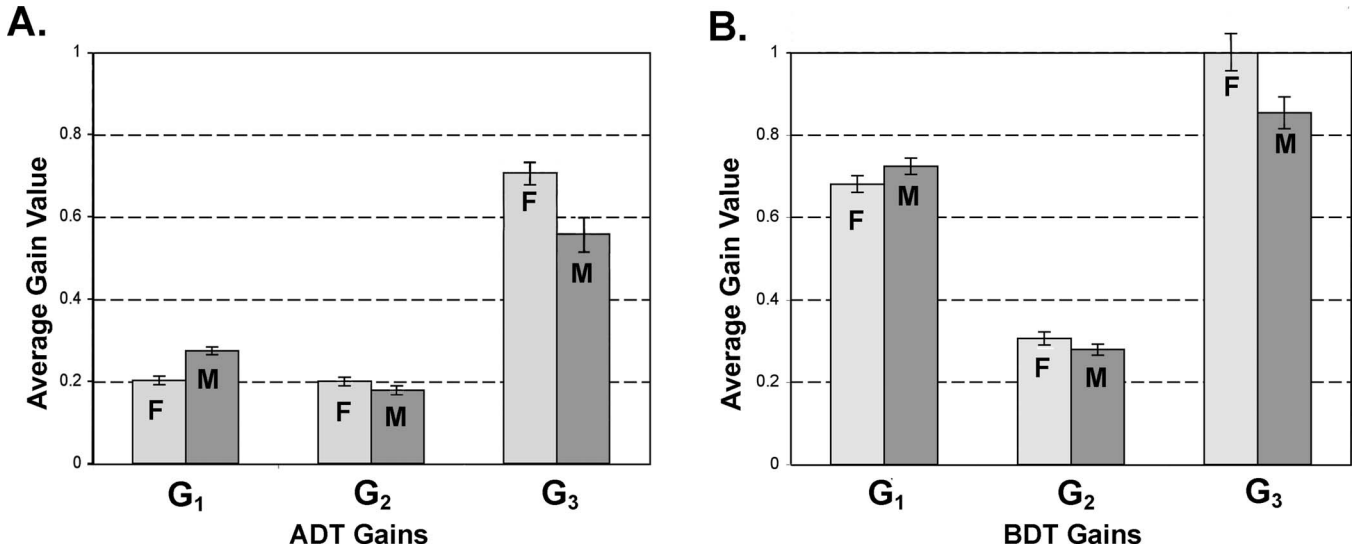


Fig. 1. Gains G_1, G_2, and G_3 by gender for astronomy and biology. The t-tests of the difference for astronomy give $p<0.0005$ for G_1, $p=0.045$ for G_2, and $p=0.004$ for G_3. For biology $p=0.123$ for G_1, $p=0.112$ for G_2, and $p=0.64$ for G_3. Error bars represent the standard error of the mean.

gain), G_2, and G_3. We found that men's average scores consistently lead to higher apparent learning gains than women's scores when G_1 was used. Conversely, the G_3 calculation resulted in consistently higher apparent learning gains for female students. There was no apparent difference in learning gains between men and women when G_2 was used (see Fig. 1). This conflicting pattern occurred for two separate instruments in both disciplines, although gender differences were more pronounced in astronomy than biology.

The fact that three measures of gain yield different learning results suggests that conclusions about gender specific learning by the examination of gain scores must be approached with care. Our data suggest that there is little, if any, verifiable difference in learning gains between men and women even if women both start and end at lower levels of achievement than men. This conclusion is bolstered by the fact that there is no identifiable difference in student performance on assignments, exams, or course grade. Instead, the differences in learning gains calculated by these methods appear to stem from their definitions.

Expressions of learning gain are used not only to correct for different levels of incoming knowledge among students but also appear to incorporate bias into the interpretation of the data.[25,35] Brogt *et al.*[25] demonstrated that pretest scores can markedly influence gain values by calculating G_1, G_2, and G_3 for three relations of pre- and post-test scores: one in which all students have identical absolute gains regardless of pre-test score, a second in which the post-test score is an incremental increase of the pretest score, and a third in which both effects occur simultaneously. In the third learning model G_1 is strongly biased toward high pretest scores, G_2 is slightly biased toward low pretest scores, and G_3 is strongly biased toward low pretest scores (see Fig. 2). The results are similar if the learning model assumes all students post an absolute gain of 25 points regardless of pretest scores. If post-test scores are strictly a function of pretest scores (post-test=1.5pretest), G_1 still skews toward high pretest scores, but the values of G_2 and G_3 are insensitive to pretest scores.

All three of the models discussed in Ref. 25 demonstrate that higher pretest scores result in disproportionately higher normalized gain values. The learning gain models that emerge from our data (post-test=0.4pretest+65 for the BDT, post-test=0.85pretest+21 for the ADT) most closely resemble the third learning model and, as shown in Fig. 1, result in the groups with higher pretest scores (males) having larger average G_1 values than the groups with lower pretest scores (females) in both courses. In contrast, the absolute gain, as well as the other gain definitions such as G_2 and G_3, suggests either no difference in learning between males and females or an opposite effect.

To further analyze if the variation in the calculated gains can be attributed to the gain definitions rather than inherent gender differences in learning, G_1, G_2, and G_3 for both courses were plotted as a function of pretest score (see Fig. 3). Although the slight positive slope of the G_1 versus pretest score regression line indicates that the normalized gain weakly favors high pretest scores (and thus provides a slight

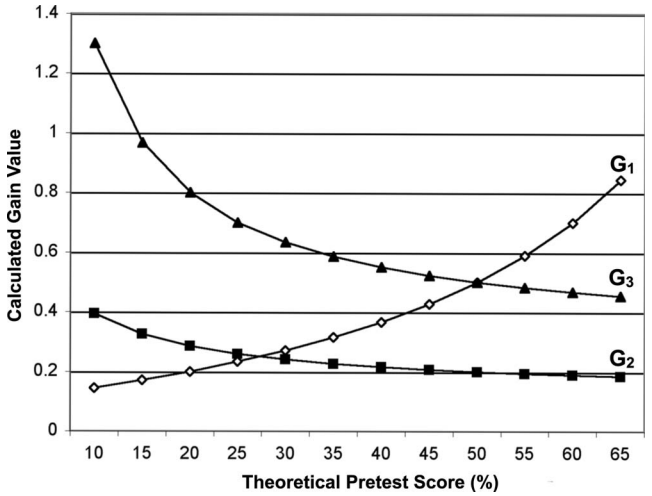


Fig. 2. Gain calculation bias for linear data models with nonzero slope and y intercept values (Ref. 25). The model assumes the relation of pretest to post-test scores as post-test=(1.3)pretest+10.

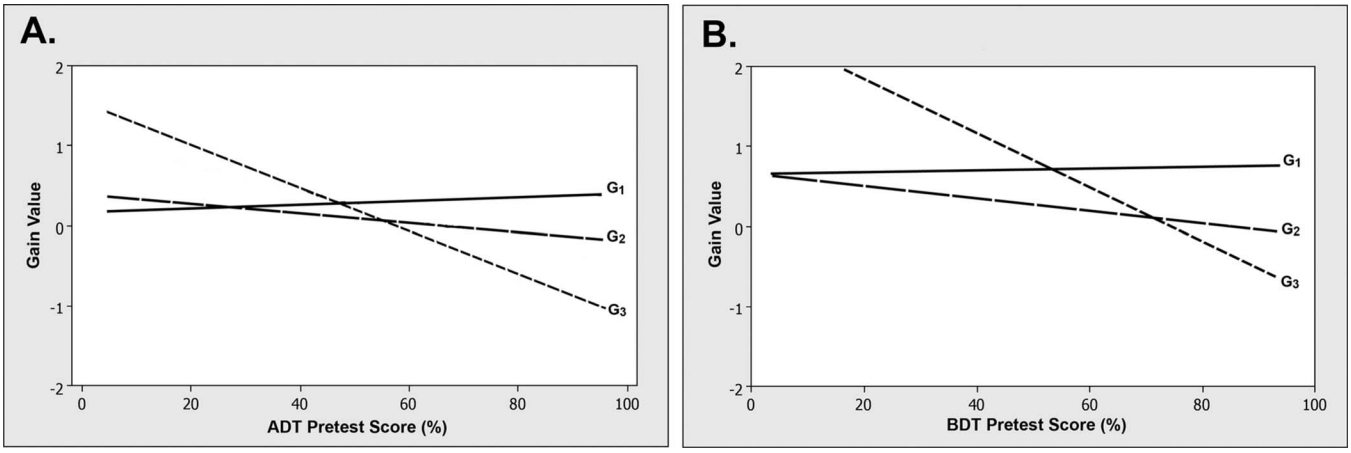


Fig. 3. Gains G_1, G_2, and G_3 regression lines as functions of pretest scores. (a) Astronomy course ($N=1133$), Pearson correlation coefficients for regression lines. $r=0.135$ ($p<0.0005$) for G_1, $r=-0.473$ ($p<0.0005$) for G_2, and $r=-0.478$ ($p<0.0005$) for G_3. (b) Biology course ($N=213$). $r=0.085$ ($p=0.215$) for G_1, $r=-0.858$ ($p<0.0005$) for G_2, and $r=-0.804$ ($p<0.0005$) for G_3.

inherent bias in calculating men's learning gains), G_3 calculates higher gain values for lower pretest scores and thus may overvalue women's gains. As can be seen in Fig. 3 the G_3 calculations were biased toward low pretest scores, which is in keeping with the findings of Ref. 25 (see Fig. 2). Note that G_3 indicates greater learning gains for women despite the fact that there is little or no difference between men and women in absolute gain between men and women (see Fig. 1 and Table II).

G_2 is least likely to show disproportionately high gains with either high or low pretest scores because the distribution of gain values calculated with this formula is symmetric about the mean. By generating a more symmetrical set of gain values, G_2 prevents the undue influence of very high or very low gain values that can result from use of either G_1 or G_3. As shown in Ref. 25, G_2 is less sensitive to low pretest scores than G_3, and less sensitive to high pretest scores than G_1 in a learning model which (like ours) has both a nonzero slope and intercept.

Given these conflicting data interpretations resulting from the use of varying gain formulas, we looked for other evidence of quantifiable differences in classroom learning between men and women. We performed a covariate analysis on the data to explore how each measure of gain depended on gender and pretest score, and to determine if there were differences between genders in learning gains that could not be accounted for by the males' higher pretest scores. The results of the covariate analysis are shown in Table III. The values of G_2 and G_3 for the ADT indicated significant interaction between the pretest and gender factors, indicating that the ANCOVA analysis is not appropriate for those data. The data suggest that although raw pretest and post-test performances vary significantly by gender, there are no significant differences in learning gains between males and females once the effect of the different pretest scores are accounted for via the analysis of variance. Multiple regression was performed on the G_2 and G_3 ADT data, and we found that there is a correlation between pretest scores, gender, and gain ($R^2=24\%$ for G_2 and $R^2=25\%$ for G_3), and hence a linear model with predictors of pretest score and gender alone might suffice for determining gains. The R^2 term is the amount of the variance in the response variable (gain) that can be accounted for by the predictor variables (pretest score and gender) and is calculated by squaring the correlation coefficient, r.

IV. DISCUSSION AND CONCLUSIONS

Pre- and post-tests have been used to measure learning gains in physics classrooms for over 20 years.[36] The normal-

Table III. Covariate analysis of gender and pretest scores on the effect of the calculated gains. Interaction terms are shown, which is the simultaneous effect that gender and pretest scores have on gains, and R^2 values indicate the variation in the response (gains) that can be accounted for by the predictors (pretest, gender, and both simultaneously). The ADT G_2 and G_3 interaction term p-values are significant, indicating that these covariate analyses may not be reliable.

Course	Gain	p-value of pretest score coefficient	p-value of gender coefficient	p-value of pretest and gender interaction term	R^2 value (%)
Astronomy	G_1	0.02	0.55	0.28	3
	G_2	<0.0005	0.88	0.03	24
	G_3	<0.0005	0.01	<0.0005	26
Biology	G_1	0.25	0.29	0.13	3
	G_2	<0.0005	0.37	0.27	74
	G_3	<0.0005	0.12	0.12	65

ized gain G_1 has been used to describe learning gains in physics and astronomy courses,[18,19,26] and in a very large study Hake[4] used G_1 to compare student gains in traditional versus interactive engagement introductory physics courses. These reports show larger learning gains for males than females, consistent with our G_1 scores.

Recently, the appropriate use of the normalized gain and other gain score calculations has been scrutinized by science education researchers. Coletta and Phillips[5] noted a strong positive correlation between pretest scores and normalized gains in their study but posited that this result might be due to hidden differences in scientific reasoning skills in different groups. Brogt *et al.*[25] showed that different expressions for calculating gains have different inherent biases, with G_1 being particularly sensitive to high pretest scores. Recognizing the possibility of bias toward high pretest scores is critical for gender analyses of learning gains because males have been shown to post significantly higher pretest scores on instruments such as the FMCE, FCI, and ADT.[3,6,18,19,27] Given that males commonly earn higher pretest scores on these instruments, combined with the fact that the normalized gain is frequently used to compare pretest and post-test scores, it is not surprising that these studies reported higher average learning gains for males compared to females.

In the current study the performance gap between males and females is present in both astronomy and biology and in both the pretest and the post-test and was evident whether the test was given online (as in the BDT pretest), in the classroom (ADT pre- and post-test), or embedded in an exam for which students had ample preparation (BDT). Because the gender disparity persists across disciplines and testing methodologies and cannot be explained by poorer overall achievement in either course by women suggests that another factor may be influencing performance.

Differences in test scores by gender may be in part due to the phenomenon of "stereotype threat"—the conscious or subconscious influence of negative stereotypes in our culture regarding (particularly women's) math and science competence.[37,38] Although females outperform males in terms of academic achievement (grades) on both stereotypically masculine and feminine subjects[39,40] and also do better on standardized tests of stereotypically feminine subjects (for example, reading and spelling),[39] they continue to underperform on standardized tests of science. (Although girls have recently achieved parity on standardized tests in math at the primary and secondary education level.)[41]

Gender-based stereotype threats occur for a woman when she fears that she will be judged by the prevailing cultural stereotype that a woman is not as good at math or science. Despite inroads at the K-12 level in math scores, evidence indicates that women continue to lag behind their male peers in math and science performance at the undergraduate level. Given that women still make up a minority of the professional ranks in the sciences, particularly in disciplines such as physics and engineering, stereotype threat provides a compelling explanation for women's consistently lower scores on standardized science tests such as the FCI and ADT. (For a review of stereotype threat in education, particularly in math and math-based disciplines, see Ref. 42.)

In a study on the effect of FCI questions that are couched in the typically male interests of hockey and rockets, an altered version of the FCI featuring stereotypically feminine contexts for questions (for instance, replacing a question about a cannonball fired off a cliff with an identical question featuring a baby girl throwing a bowl from a highchair) was administered to 300 college students.[20] The performance of men and women on this alternative FCI was compared to the performance of 300 men and women on the unaltered FCI at the same institution. Although the average overall score for women did not statistically improve, the men's average score decreased significantly. Changes in the student performance on individual items of the altered FCI were unpredictable. Both women's and men's scores on revised questions sometimes improved, sometimes worsened, and sometimes stayed the same (depending on the question) and this variability suggests that the cultural references used in phrasing questions on assessment instruments do affect student performance. However, determining how cultural references alone effect student performance is difficult if not impossible in a real-world setting.

Stereotype threat appears to be a plausible explanation for gender gaps in standardized test performance, but it is problematic to assume that it is the sole explanation for differences in diagnostic test performance across gender in our classrooms. If stereotype threat were present in our classrooms, there should be a measurable difference between genders on all exam scores (not just diagnostic tests) and in overall course grades as well. Analysis of test scores in both the biology and astronomy courses found no significant difference in overall exam performance across gender, and no significant difference in earned course grades, consistent with another (smaller) study that also found no evidence of stereotype threat in lower level astronomy courses.[43] The gender gap in assessment instrument performance in our courses may be due to a more complex interaction of factors which may include incoming knowledge, stereotype threat, or other more subtle cultural differences.

Although performance gaps between males and females on all manner of science diagnostic tests are well documented and continue to be observed, we must pay particular attention to arguments regarding the ability of women to learn (or not learn) as effectively as their male peers in university science courses. Differences in classroom learning gains (particularly as determined using the normalized gain calculation) seem to indicate that men are better learners. Such a conclusion is not borne out by analyzing the data using other calculation methods or by examining test scores, laboratory/homework scores, and in-class activity scores. Our analysis of learning gains in biology and astronomy classrooms shows no evidence that women are less capable of learning than their male peers. Given the biases introduced by various learning gains, we suggest caution when using these measures to draw conclusions about differences in science classroom performance across gender.

ACKNOWLEDGMENTS

The authors would like to thank Chuck Paden and Cali Morrison for assisting with the development of the BDT.

[a]Electronic mail: willoughby@physics.montana.edu

[1]L. C. McDermott and E. F. Redish, "Resource Letter: PER-1: Physics education research," Am. J. Phys. **67**, 755–767 (1999).
[2]D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).
[3]G. L. Deming, "Results of the astronomy diagnostic test national project," Astron. Educ. Rev. **1**, 52–57 (2001).

[4] 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**, 64–74 (1998).
[5] V. P. Coletta and J. A. Phillips, "Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability," Am. J. Phys. **73**, 1172–1182 (2005).
[6] M. K. Hemenway, W. J. Straits, R. R. Wilke, and B. Hufnagel, "Educational research in an introductory astronomy course," Innovative High. Educ. **26**, 271–280 (2002).
[7] B. Hufnagel, T. Slater, G. L. Deming, J. Adams, R. L. Adrian, C. Brick, and M. Zeilik, "Pre-course results from the astronomy diagnostic test," Electr. Publ. Astron. Soc. Aust. **17**(2), 152–155 (2000).
[8] Information available at ⟨http://www.bioliteracy.net⟩.
[9] K. Garvin-Doxas, M. Klymkowsky, and S. Elrod, "Building, using and maximizing the impact of concept inventories in the biological sciences: Report on the National Science Foundation-sponsored conference on the construction of concept inventories in the biological sciences," CBE Life Sci. Educ. **6**, 277–282 (2007).
[10] K. Garvin-Doxas and M. Klymkowsky, "Building the biology concept inventory," ⟨http://www.bioliteracy.net/Readings/papersSubmittedPDF/Gavin-Doxas%20and%20Klymkowsky.pdf⟩.
[11] E. Seymour, D. J. Wiese, and A. B. Hunter, "Creating a better mousetrap: On-line student assessment of their learning gains," Paper presented at the *National Meeting of the American Chemical Society*, San Francisco, March 2000 (http://www.sencer.net/Assessmemt/pdfs/Assessment/Creating_a_Better_Mousetrap.pdf).
[12] N. Bowers, M. Brandon, and C. D. Hill, "The use of a knowledge survey as an indicator of student learning in an introductory biology course," CBE Life Sci. Educ. **4**, 311–322 (2005).
[13] *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*, edited by B. S. Bloom (Longman, New York, 1956).
[14] Educational Testing Service, ⟨http://www.ets.org⟩.
[15] G. Z. Wilder and K. Powell, "Sex differences in test performance: A survey of the literature," *College Board Report 89–3 Sex Differences in Test Performance: A Survey of the Literature* (College Entrance Examination Board, New York, 1989).
[16] N. Cole, *The ETS Gender Study: How Females and Males Perform in Educational Settings* (Educational Testing Service, Princeton, NJ, 1997).
[17] R. Thornton and D. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," Am. J. Phys. **66**, 338–352 (1998).
[18] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, "Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?," Phys. Rev. ST: Phys. Educ. Res. **3**, 010107–010110 (2007).
[19] M. Lorenzo, C. H. Crouch, and E. Mazur, "Reducing the gender gap in the physics classroom," Am. J. Phys. **74**, 118–122 (2006).
[20] L. McCullough, "Gender, context, and physics assessment," J. Int. Women's Studies **5**, 20–30 (2004).
[21] Summary data for April/August 2005 MCAT, ⟨http://www.aamc.org/students/mcat/examineedata/pubs.htm⟩.
[22] Summary data for April/August 1991 MCAT, ⟨http://www.aamc.org/students/mcat/examineedata/pubs.htm⟩.
[23] D. Ritter, "Gender role orientation and performance of stereotypically feminine and masculine cognitive tasks," Sex Roles: A Journal of Research **50**, 583–591 (2004).
[24] G. E. Francis, J. P. Adams, and E. J. Noonan, "Do they stay fixed?," Phys. Teach. **36**, 488–490 (1998).
[25] 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**, 25–42 (2007).
[26] M. Zeilik and V. J. Morris, "An examination of misconceptions in an astronomy course for science, mathematics, and engineering majors," Astron. Educ. Rev. **2**, 101–119 (2003).
[27] R. R. Hake, "Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization," online at http://www.physics.indiana.edu/~hake/PERC2002h-Hake.pdf.
[28] See EPAPS Document No. E-AJPIAS-74-003603 for average gains calculated from Ref. 19. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[29] The BCI was unavailable at the start of this study and also has a heavy emphasis on genetics and evolution that did not match well with the course content of the biology course examined in this study.
[30] R. M. Kaplan and D. P. Saccuzzo, *Psychological Testing: Principles, Application and Issues* (Wadsworth, Belmont, CA, 1997).
[31] D. Hestenes and I. Halloun, "Interpreting the Force Concept Inventory," Phys. Teach. **33**, 502–506 (1995).
[32] Individual student's pretest scores, post-test scores, and gender were linked for both courses, and identifying information was removed from the data set prior to analysis. The study in both courses was performed with the review and approval of the Institutional Review Board for the Protection of Human Subjects at MSU.
[33] L. Turnquist, P. Vatria, and Y. Vartia, "How should relative changes be measured?," Am. Stat. **39**, 43–46 (1985).
[34] L. Bao, "Theoretical comparisons of average normalized gain calculations," Am. J. Phys. **74**, 917–922 (2006).
[35] J. D. Marx and K. Cummings, "Normalized change," Am. J. Phys. **75**, 87–91 (2007).
[36] I. Halloun and D. Hestenes, "Common sense concepts about motion," Am. J. Phys. **53**, 1056–1065 (1985).
[37] C. M. Steele, S. J. Spencer, and J. Aronson, in *Advances in Experimental Social Psychology*, edited by M. Zanna (Academic, New York, 2002), p. 379.
[38] G. A. Kenney-Benson, E. M. Pomerantz, A. M. Ryan, and H. Patrick, "Sex differences in math performance: The role of children's approach to schoolwork," Dev. Psychol. **42**, 11–26 (2006).
[39] *Gender Gaps: Where Schools Still Fail our Cjildren* (American Association of University Women Educational Foundation, Washington, D.C., 1998).
[40] E. M. Pomerantz, E. R. Altermatt, and J. L. Saxon, "Making the grade but feeling distressed: Gender difference in academic performance and internal distress," J. Educ. Psychol. **94**, 396–404 (2002).
[41] "The nation's report card: Mathematics highlights 2004," Report No. NCES2004-451, 2004.
[42] C. S. Smith and L. Hung, "Stereotype threat: Effects on education," Soc. Psychol. Educ. **11**, 243–257 (2008).
[43] B. Hufnagel, G. Deming, J. Landato, and A. Hodari, "The effect of stereotype threat on undergraduates in an introductory astronomy class," J. Women Minor. Sci. Eng. **10**, 89–98 (2004).