The Performance gender gap in undergraduate physics

Ms Robyn Donnelly, Dr Cait MacPhee & Prof Simon Bates
School of Physics and Astronomy
University of Edinburgh, Edinburgh, EH9 3JZ
R.C.A.Donnelly@sms.ed.ac.uk cait.macphee@ed.ac.uk
s.p.bates@ed.ac.uk

Abstract

This paper presents results from a study investigating the existence and degree of gender disparity in various types of formative and summative assessment in a first year undergraduate physics course at the University of Edinburgh. Differences in performance on standard diagnostic tests of conceptual understanding, continuous assessment, formative in-class Peer Instruction discussions and examination results of male and female sub-groups of the cohort have been investigated. We have found that male students are significantly outperforming female students on conceptual understanding diagnostic tests taken at the point of entry to university and that this performance gender gap persists after a semester of teaching. Female students outperform males on continually assessed work, while for end-of-course examinations the picture is less clear. There is no significant gender dependence on average learning gains seen during Peer Instruction episodes in lectures, though individual question pairs can display marked gender differences.

Keywords

Physics, gender, clickers, peer instruction

1. Introduction

The under-representation of women at all stages of the student lifecycle, with regard to physics, is of increasing concern and has often been described as the “leaky pipeline” (Blickenstaff, 2005). Results released by the Institute of Physics (IOP) indicate that the proportion of females entering first year undergraduate physics has remained consistently around 18-20% for the past 15 years (IOP, 2012). This compares with much higher levels of female participation in mathematics (41%) and chemistry (48%) (JCQ, 2012). In addition to participation levels there is indication that a potential gender disparity in performance in undergraduate education is present (Lorenzo et al, 2006). It has been argued that differences of background preparation in physics and mathematics of incoming students may explain the evident gender disparities in participation and performance at university level.

This provides additional motivation to consider the effectiveness of instructional methodologies within STEM disciplines; both to ensure effective learning for the entire cohort and to investigate if there is a performance gender gap. This study investigates student learning on different types of assessment: those created specifically for students taking the particular courses we offer (course assessments, end of course examinations etc) and also standardized diagnostic instruments, designed to probe conceptual understanding in the subject matter. Use of such standardized instruments permits meaningful comparison with other courses, taken by other students at other institutions.

One such instrument is the Force Concept Inventory (FCI) (Hestenes and Wells, 1992). The FCI has been extensively used worldwide as a diagnostic tool to measure students' understanding of Newtonian concepts of force and highlight common misconceptions. The test consists of 30 questions probing six individual concepts relating to Newtonian mechanics. A score of 60% on the FCI is classified as an "entry threshold" for understanding of Newtonian mechanics and a score of 85% as a "mastery threshold" (Hestenes and Halloun, 1995). It is often used as an instrument to measure students' understanding at the point of entry to university or the beginning of an introductory level physics course (Pre-test). Repeating this test at the end of a course (Post-test) allows for it to be used as a measure of the effect a course has on students' understanding of the tested concepts.

There is growing evidence to suggest that specific teaching methodologies and pedagogical approach both have a measurable effect on students' overall performance and can increase learning more than traditional methods (Hake, 1998). Hake's analysis of the normalized gains (change in score from pre-test to post-test as a fraction of total possible increase in score) on pre- and post-instruction test scores on the FCI of nearly 6000 students at US institutions has provided a benchmark for comparison of effective learning and teaching methodologies. Hake's study found that the learning gains (measured using the FCI) on courses taught to facilitate "interactive engagement" (IE) of students were approximately twice those in traditionally taught courses. A key component in many (but not all) IE courses was the use of clickers in lectures, to stimulate discussion and debate, with clicker episodes built around the Peer Instruction method (Mazur, 1997). The differences in implementation of Peer Instruction (PI) in lectures, and the subsequent effect on student learning gains, has also been discussed extensively (Turpen and Finkelstein, 2009).

Previous research by Docktor and Heller using the FCI indicated that females enter university with a significantly lower understanding of Newtonian concepts (Docktor and Heller, 2008), with females on average scoring 15.3±0.5% less than males. This statistically significant gender gap persisted after one semester of teaching, with both cohorts exhibiting a similar gain. A similar study by Lorenzo et al. (Lorenzo et al, 2006) in an introductory calculus-based physics course at Harvard University found an increase in both male and female learning gains through the instigation of interactive engagement techniques and PI over the course of a semester. In this study, the pre-semester gender gap was fully closed at the end of the course, which the authors attribute to the PI methodolgy. This is not, however, a result that has been universally reproduced and there is suggestion that both instructor effects and background experience of students may be factors in explaining observed differences (Pollock et al, 2007).

Looking beyond standardized diagnostic tests, evidence has suggested that different genders perform differently depending on type of assessment. Research by the University of Colorado analysed gender difference in students' coursework and examination grades to investigate potential gender bias in assessment (Kost et al, 2009). Although there was no apparent gender discrepancy in overall course grade, in each of the seven semesters tested males consistently outperformed females on exams, whilst females scored consistently higher than males on coursework. In an earlier study, the authors note that coursework assignments are designed to be collaborative with little time dependence, whilst exams are individual and involve an element of competition with a specific time constraint, which may favour particular students or possibly genders.

This study looks at student performance as a function of gender on a variety of assessments in our first year physics courses at the University of Edinburgh. This course has progressively incorporated more and more elements of an IE methodology over the last 6 years, including use of clickers and PI. We consider three distinct assessment scenarios: pre- and post-instruction performance on the FCI; summative coursework and examinations; and formative in-lecture PI episodes. The first two of these are replication studies in a different educational environment; the latter has to our knowledge not been undertaken or reported previously.

2. Results

"Physics 1A: Foundations" is a first-year, first-semester course in classical mechanics and dynamics. The course is compulsory for all students registered on a physics degree programme (approximately 50% of the class) and is also a popular elective with students registered on other programmes. Class sizes have varied significantly in recent years, ranging from a minimum of 199 to a maximum of 320. The proportion of female students enrolled in Physics 1A has remained between 20-29% between 2006-2011. In each section, data from the 2011/12 cohort will be discussed separately due to both changes in recruitment and selection policies for incoming students in this year, and also the structure of the Physics 1A course.

2.1 Force concept inventory

Between 2006-2010 a revised and expanded version of the FCI survey, containing 33 questions (Halloun, 2012), has been employed at the University of Edinburgh[1]. Mean scores for each year were sufficiently similar that we were able to combine the five individual years of data. Results from this composite matched data set of 858 students (ie students who have taken both the pre- and post- tests) demonstrate the existence of a statistically significant gender gap ($p<0.05$) consistent with that seen in other studies in the literature. On entering university, prior to any instruction, males score on average 12% higher on the FCI than incoming females. Post-test results were collected in week 8 of the semester (at this stage all the material within the FCI had been covered) and illustrated that this gap decreases greatly but nevertheless remains significant with males now outperforming females by only 4%. Consistent with the narrowing of the gender gap[2] is the fact that female students have a

[1] Although when introduced to the course in 2006 the diagnostic test was not embedded fully in the curriculum as a weekly assignment as in following years.

[2] We have defined the gender gap as male score minus female score. For each of the five years in question the change in the gender gap (gap post-test minus gap pre-test) was negative, indicating a consistent move towards closing (or narrowing) of the gender gap after one semester of teaching.

higher normalised gain than male students following one semester of teaching; 0.53 for female students and 0.49 for male students.

When data was examined on a question-by-question basis, a gender gap on some questions was also evident (though there was a large variation in percentage of correct responses, indicating greater or less understanding of certain physical concepts). For almost all questions there was a lower percentage of correct responses for females compared to males with the majority of questions (25 out of 33) showing a statistically significant difference. As in the overall results, the gender gap was significantly reduced post-test with the number of questions showing a significant gap decreasing to 14.

At the start of the 2011 session, the original version of the FCI containing 30 questions was used and a similar gender dependence was seen compared to previous years. In the pre-test results male students (N=116) scored on average 10.6% higher than female students (N=45). This gender gap reduced to 7.7% post-test but again the gap remained significant (p=0.012) (Figure 1a). If we consider the normalised gains associated with this test instrument, we find once again both genders achieved high normalised gains: 0.57 and 050 for males and females, respectively.

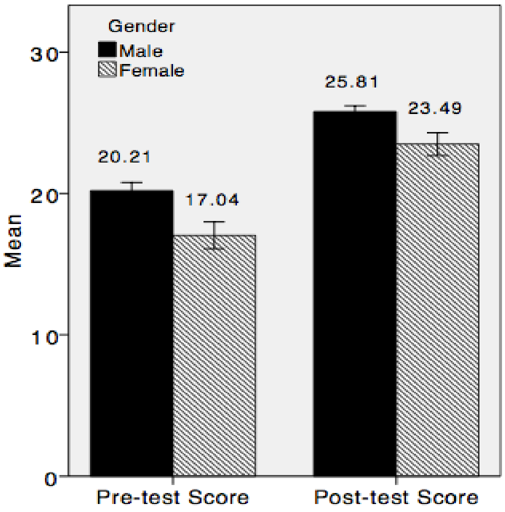

Figure 1a

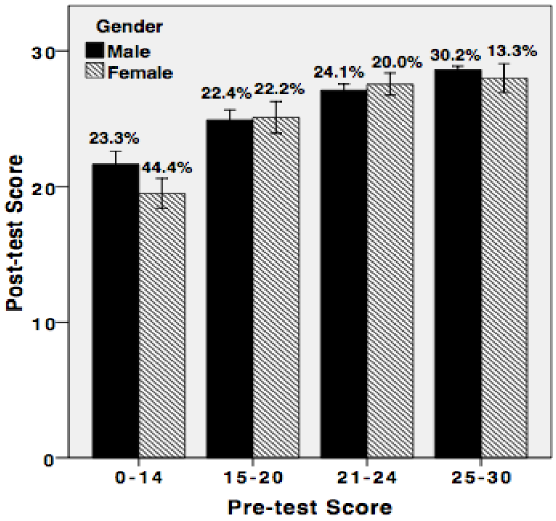

Figure 1b

Figure 1a): FCI pre- and post-instruction scores as a function of gender for the 2011/12 cohort. Error bars represent the standard error of the mean. Values above each bar represent mean score out of 30. Figure 1b): 2011/12 FCI scores binned as a function of pre-instruction test score. Error bars represent the standard error of the mean. Values above each bar represent percentage of each individual cohort represented in each of the four quartiles.

Preliminary analysis of 2011 data has been performed to investigate if there is a gender difference in post-instruction test score as a function of pre-instruction score. Students who completed both the pre-test and post-test were binned into four cohorts of approximately equal population, dependent on their FCI pre-test scores and the average post-test score calculated as a function of gender in each bin. The results are shown in Figure 1b and although there is no statistically significant gender gap in any of the four bins, there are several striking features. The bottom quartile, on the basis of pre-instruction test score, comprises almost half (44%) of the entire female population in the class. This is significantly larger than the male population in this quartile (23%). Indeed, males are distributed rather evenly through the pre-instruction score distribution, whereas female students are concentrated towards the lower quartiles, with only 13% of all females in the top quartile group. The largest discrepancy in mean post-instruction score is seen in the lowest quartile (though this is not significant at the 95% confidence level). It does however appear that the majority of the post-instruction gender gap arises from performance of the weaker students in the cohort (which include a proportionately larger fraction of female students). Analysis of historical data for previous years will enable to address if this is a consistent feature seen across multiple years.

2.2 Coursework and exam assessment

In addition to student performance on conceptual tests, we consider in-course assessments and end-of-course examinations. Analysis of results showed that females consistently outperformed males in coursework in each year between 2006 and 2010. If we combined the data, the average difference is females outperform males by 7% and this is a statistically significant result ($p<0.01$) (Figure 2). Although the pattern of performance on coursework is consistent with results from the University of Colorado study (Kost et al, 2009), the exam scores did not demonstrate as distinct a pattern. In the examination scores, there is far more variation between individual year groups, with females outperforming males in two of the five years. This is perhaps not wholly unexpected since, in almost all cases, the questions used in coursework exercises remain the same each year but exam questions are changed on a yearly basis. We would therefore expect greater variation in exam performance year-on-year.

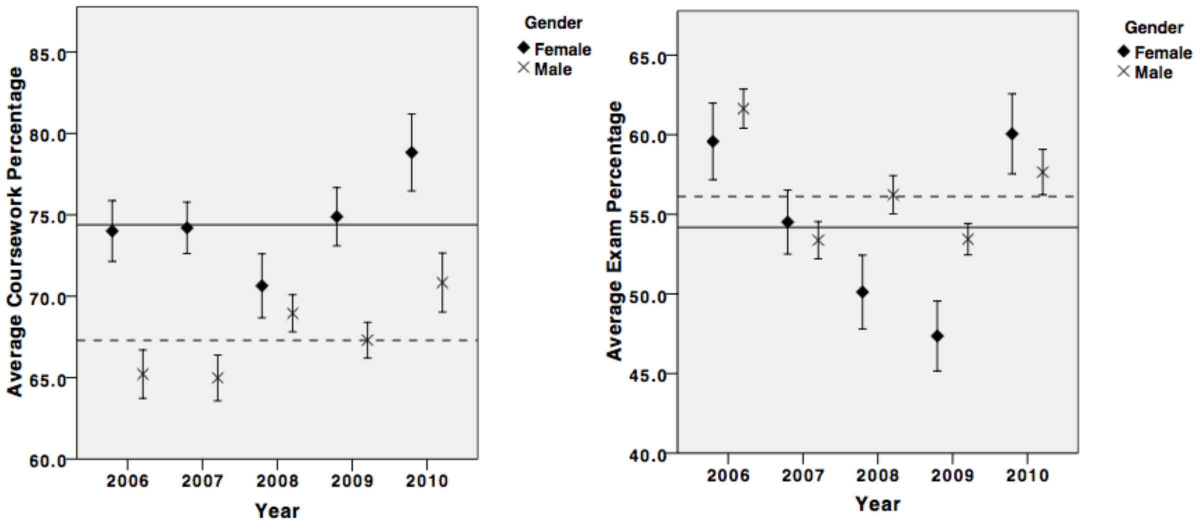

Figure 2: Average coursework and exam scores for Physics 1A students between 2006-2010 as a function of gender. Error bars represent the standard error on the mean. Horizontal lines represent average scores for male (dashed) and female (solid) students.

In 2011 the structure of the first semester physics course changed extensively with students receiving the content material of the course prior to lectures with the lectures themselves focusing on PI (Bates and Galloway, 2012). The end of course examination was open-book. Nevertheless, a similar pattern to previous years is seen with male students scoring slightly lower than females, although not significantly so. Males on average scored 64.6% compared to females 66.7% in the overall coursework element of the course. The end of course exam scores showed no gender difference with both genders scoring within 0.1% of each other.

These findings provide motivation for a wider study of the extent to which instructional methodologies affect a performance gender gap and whether this apparent gender gap in coursework assessment persists through second year and into honours physics courses or whether it is characteristic of the introductory physics course.

2.3 Clickers

The Physics 1A lectures made extensive use of PI. This instructional technique involves students being asked questions testing their conceptual understanding of the topic being studied, and answers collected via electronic voting handsets or clickers. Where the proportion of correct individual responses is between 30-70%, students were encouraged to discuss the question and possible answers with neighbouring students, to state and defend their views, before a second round of voting took place. This is then followed by discussion or explanation led by the course lecturer. We loan students a clicker handset for the entire semester, which enables us to assign handset id's to students for post-lecture analysis.

A sample of that analysis is presented here, for clicker questions administered in the first 5 weeks of teaching. The content of this section of the course covered Newtonian mechanics concepts dealt with in the FCI. In total in this section of the course, 65 clicker questions were asked, of which 14 involved full PI discussion (vote and revote). The participation rate for each questions ranged from 56% to 77% of students registered on the course.

A data set containing matched student data[3] was created. Analysis of the 14 PI question pairs showed that males slightly outperformed females in average pre-discussion percentage correct responses although this difference was not significant and varied considerably depending on individual questions. Normalised gains[4] for paired questions suggest an overall positive effect from peer discussion. An average gain of 0.5 was achieved by both genders. All but two questions had a normalised gain over 0.3, with some questions exhibiting gains of 0.9.

Despite there being no significant gender difference in students' overall performance on PI paired questions, individual 'clicker' questions showed signs of gender differences in performance, with males having a much higher percentage correct response rate compared to females in specific questions and a greater percentage of females answering correctly compared to males on others.

[3] Students entering a response to both the initial vote and post-discussion vote.

[4] Where normalised gain has been defined as the percentage of students improving post discussion as a fraction of the percentage of students who initially voted incorrectly pre discussion.

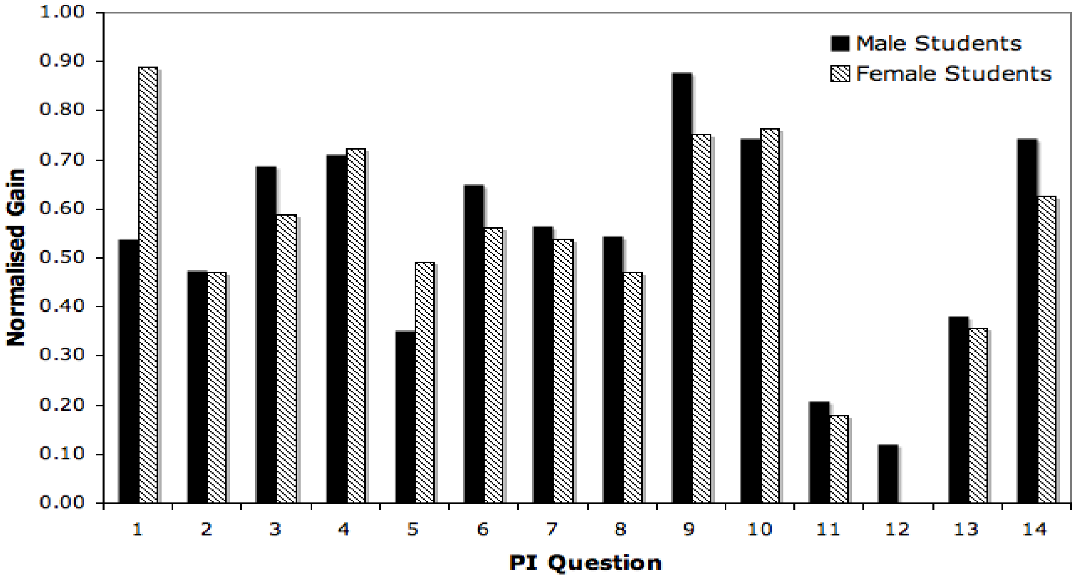

Figure 3: Normalised Gains for PI questions as a function of gender

This is seen in Figure 3 which shows normalised gains for the 14 PI questions as a function of gender. On 2 of the 14 PI questions, females have a normalised gain of at least 0.02 greater than the male cohort. On PI question 1, a question essentially concerned with rounding to the correct number of significant figures, females dramatically outperform males on the re-vote. Percentage correct scores for females exceeded 93% (a normalised gain of 0.89) compared to 74% for males (a normalised gain of 0.54). Both genders had very low normalised gains on question 12. This question is an example of unsuccessful PI episode, with males achieving on average a normalised gain of 0.12 and females a normalised gain of zero. In this case there was very little change in answer between the vote and revote, and only 8 out of 31 female students changed their vote after PI discussion. We intend to extend this analysis across all PI questions in the course to investigate the existence of any trends in gender performance on a question by question basis.

3. Discussion

The results presented in this paper indicate that as well as the existing participation gender gap, a significant performance gap is present between male and female students at the point of entry into undergraduate physics courses. The administration of annual conceptual tests to incoming physics students has indicated that this differential performance by sub-cohort groups is seen every year and that the significant difference between genders remains after one semester of teaching, but is narrowed.

The observed difference in trends in coursework and examination suggests a possible inclination of females towards continually assessed coursework components of the course offers the prospect for further investigation into the potential gender preference in assessment styles. In order to further understand the underlying source of these gender issues we aim, through future study, to determine whether this apparent gender difference in coursework assessment continues as students progress through their university degree and whether similar trends are witnessed in other STEM subjects at the University of Edinburgh which display very different gender profiles (specifically chemistry and biology).

Results from a preliminary analysis of in-lecture PI questions have suggested an equally beneficial effect on student learning and engagement in lectures for both male and female students. Learning gains on individual PI episodes for both genders are comparable to those seen overall on the FCI. There is some suggestion that there may be a gender element to performance on a question by question basis. This also requires further exploration to determine whether this is assessment or question dependent.

4. References

Bates, S. and Galloway, R., (2012) (in press). *The inverted classroom in a large enrolment introductory Physics course: a case study.* Proceedings of the 1st International Conference on the Aiming for Excellence in STEM Learning and Teaching

Blickenstaff, J.C., (2005) *Women and science careers: leaky pipeline or gender filter?* Gender and Education, 17(4), 369-386.

Docktor J., Heller K. (2008) *Gender differences in both force concept inventory and introductory physics performance.* AIP Conference Proceedings, 1064(1), 15-18

Hake, R. (1998) *Interactive-engagement vs traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses,* Am. J. Phys., 66(1), 64-74

Halloun,I. Assessment: Inventory of Basic Concepts in Mechanics. Available from http://www.halloun.net/index.php?option=com_content&task=view&id=4&Itemid=6, [accessed 20/02/2012]

Hestenes D., Halloun I. (1995) *Interpreting the force concept inventory.* The Physics Teacher, 33, 502-506

Hestenes D., Wells M., Swackhamer G., (1992) *Force Concept Inventory*, The Physics Teacher, 30, 141-158
Institute of Physics (IOP), http://www.iop.org/policy/statistics/education/page_43181.html [accessed 05/02/2012]

Joint Council for Qualifications (JCQ) examination results summer 2010, http://www.jcq.org.uk/national_results /index.cfm [accessed 05/02/2012]

Kost, L.E., Pollock S.J., Finkelstein N.D., (2009) *Characterizing the gender gap in introductory physics*, Phys. Rev. ST Phys. Educ. Res., 5(1)

Lorenzo, M., Crouch C.H., Mazur E. (2006) *Reducing the gender gap in the physics classroom.* Am. J. Phys. 74(2), 118-122

Mazur, E., (1997). *Peer Instruction: A user's manual. Series in educational innovation.* Prentice Hall.

Pollock, S., Finkelstein N., Kost L. (2007) *Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?* Phys. Rev. ST Phys. Educ. Res., 3 (1):010107

Turpen C., Finkelstein, N. (2009) *Not all interactive engagement is the same: Variations in physics professors' implementation of Peer Instruction.* Phys. Rev. ST Phys. Educ. Res., 5, 020101.