eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Gender differences in the Force Concept Inventory for different educational levels in the United Kingdom

Matthew Mears[*]

Department of Physics and Astronomy, The University of Sheffield, Sheffield, S3 7RH, United Kingdom

The Force Concept Inventory (FCI) is widely used to investigate the effect of education level on conceptual understanding of Newtonian mechanics but has only recently been scrutinized for gender effects and retention. This study examines both the gender gap in first year physics undergraduates compared to the gap for nonphysicists and the FCI retention after three months. All participants were either studying or working at the University of Sheffield in the UK and had completed a similar compulsory level of secondary education. As expected the results show that a greater level of education in physics is associated with a larger average FCI score. However, further analysis shows that there exists a gender gap at all levels of education. The size of the effect of gender is quantified using Cohen's d and ranges from 0.84 to 1.17 which indicates a large effect due to gender for all levels of education. Despite the FCI having been used as a tool to measure learning gains immediately following instruction in Newtonian mechanics there has been little work to investigate whether this increase in FCI score remains after some time has elapsed. Here the increase in FCI scores is found to remain increased after a three month absence of mechanics-related teaching, and that this retention of FCI scores is independent of gender. Despite this, the gender gap still remains large and statistically significant after the three month delay.

I. INTRODUCTION

Educators and policy makers have long been concerned by the underrepresentation of women in science, technology, engineering, and mathematics (STEM) disciplines in both higher education and professional careers [1,2]. In the U.S. only 25% of STEM bachelor's degrees and 26.3% of Ph.D.s were awarded to women in 2011 [3]. A similar trend is also found in the UK [4]. This representation remains into employment with only 24% of STEM workers being female (based on both 2000 and 2009 census data in the U.S. [5] and a similar proportion in the UK [6]). The potential underlying causes for this gender gap range from biological factors such as pre- and postnatal exposure to hormones [7] to lower self esteem during introductory science and mathematics courses being factors behind women leaving science and engineering majors [8]. For more in-depth discussion of the research into the underlying causes of the gender gap the interested reader is directed to the work of Halpern and colleagues in 2007 [9] and a complementary and updated review by Ceci and

colleagues in 2014 [3]. However, despite there being no consensus on the factors underpinning the gender gap, a number of strategies have been implemented to narrow it. These include the use of everyday experiences relevant to both males and females [10,11], alternating between group discussion and structured teaching [12], the use of active pedagogies [10,13,14], and providing a more diverse range of assessment and feedback methods [15,16].

The role of assessment is critical for teaching physics at college and university through problem solving [17–21] but the need for practice and the development of effective strategies [19] mean that many students completing an introductory course in physics only develop a weak understanding of the underlying concepts [22–26]. There are a number of systems for evaluating conceptual understanding of mechanics in physics [25,27,28] but the most well known and widely used is the Force Concept Inventory (FCI) developed by Hestenes, Wells, and Swackhamer [24] based on their earlier mechanics baseline test [29].

Since their initial work a number of other researchers have used the FCI to measure potential learning gains following instruction in an introductory physics course. The first large scale study ($N = 6542$) was conducted by Hake [30], who defined the normalized learning gain as

$$\langle\langle g \rangle\rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100 - \langle S_i \rangle}, \tag{1}$$

[*]m.mears@sheffield.ac.uk

where $\langle S_i \rangle$ and $\langle S_f \rangle$ are the initial and final average percentage scores, respectively. This study showed that courses taught using traditional methods ($N_T = 14$) (such as didactic lectures, recipe-driven laboratories, and algorithmic-problem exams) led to low positive learning gains ($\langle\!\langle\!\langle g \rangle\!\rangle\!\rangle_T = 0.23 \pm 0.04$) whereas 85% of "interactive engagement" courses ($N_{IE} = 41$) showed medium learning gains ($\langle\!\langle\!\langle g \rangle\!\rangle\!\rangle_{IE} = 0.48 \pm 0.14$). Similar large-scale studies have demonstrated comparable normalized learning gains as measured using the FCI [31], and there is some evidence that these learning gains persist for a number of years after education [32].

The FCI has not been without its criticisms. Huffman and Heller [33,34] conducted a factor analysis of the data presented by Hestenes and colleagues in their original work and concluded that it is unclear whether the FCI tests an understanding of key concepts or whether students achieve high scores through "…small bits and pieces of knowledge…" or familiarity with the context of the question. A more recent factor analysis [35] suggests that a five-factor model accounts for 40% of the variation in their sample ($N = 2109$) but that some questions in the FCI conflate multiple factors, most notably Newton's first and third laws, and that the kinematics concepts originally listed by Hestenes and colleagues do not appear to form a distinct factor following factor analysis. Both studies argue that even though the FCI may be a useful diagnostic tool they advise that it should not be used as a summative assessment given that there is still some uncertainty as to what underlying constructs the FCI actually measures.

With these criticisms in mind the FCI has been subsequently used in a number of different ways. Savinainen and Scott [36] demonstrate how this diagnostic tool can be used to shape and refocus teaching activities, whereas a number of researchers in the U.S. [37–39] and UK [40] have used the FCI to explore the participation and performance gap between males and females studying physics in higher education. This gender gap has also been observed across the world and at both secondary and postsecondary levels, [10,14,30,41] with the average score of females being consistently lower than males.

In this study the gender gap in first year physics undergraduates is compared to the gap for nonphysicists, all studying or working at the University of Sheffield in the UK. The retention of increased FCI scores is also examined following the U.S.-based work of Francis, Adams, and Noonan [32], however, here the cohort is based in the UK and a retention gender gap is investigated.

II. METHOD

A. Sample demographic

A significant number of studies examining the FCI are based in the United States whereas the participants in this study are from a different educational background and environment. In the United Kingdom all children are required to attend school up to the age of 16 at which point they sit for a set of national General Certificate of Secondary Education (GCSE) qualification examinations. Participants who sat their mandatory qualifications at 16 before 1988 would have sat a different but equivalent qualification known as ordinary level or "O level."

For the participants involved at age 16 they could either leave education or continue by taking further qualifications known as A levels (advanced levels), marked on a letter-based scale A*–E. These are typically seen as the entrance qualification for university study, with three qualifications being a common entry requirement for most courses. Students who wish to study physics at the University of Sheffield need a minimum of one A and two B grades which must include physics and mathematics although the majority exceeded this with an average grade profile of AAB. There are also distinct differences between UK and U.S. university courses. Students studying a course in the UK have already selected their degree course at admission and will spend their time only studying their chosen subject (or subjects in the case of dual degrees).

In this study two participant groups were recruited. Group A ($N = 125$) was recruited via a volunteer email distributed to all staff and students with an invitation to take part in an online questionnaire. The invitation email stated that anyone who is currently studying for or already holds a degree in physics should not take part. The data were checked for any respondents stating a degree or higher in physics and any such data were excluded from analysis. Group B was recruited from a class of 174 undergraduate students in their first year of a physics degree, all meeting the minimum A-level qualifications described previously.

B. Data collection

A revised version of the original FCI, known as the FCI v95, was used in this study. Furthermore the recent work by Traxler and colleagues [42] demonstrated that certain questions in the FCI v95 displayed a "…gender-unfairness…" that give rise to an inflated false gender gap, although they do note that a gender gap persists even after they removed the gender-unfair questions from analysis. In this study the "gender-unfair" questions highlighted by Traxler and colleagues were excluded from analysis but were present when the FCI was administered. Items 6, 9, 12, 14, 15, 22, 23, and 27 were highlighted by Traxler and colleagues based on their work combined with previous studies, and they also identified items 21, 24, and 29 based on their work alone.

1. Group A

For group A the reduced FCI v95 test was reconstructed into a Google form in the following structure. First, participants were provided with information about the study and to give required consent. Next the FCI questions

were presented with only one question on screen at any one time. Finally, participants were asked to complete a set of optional demographic questions including gender identity (select from "female," "male," or "prefer to self define"), age, highest qualification in physics, highest qualification other than physics, and department or professional services team (free text).

Group A is subsequently split into two groups: the "GCSE or O level" group whose highest qualification in physics is either GCSE or O level, and the "A level" group who have an A level in physics. The GCSE or O-level group was made up of 75 participants ($n_{female} = 47, n_{male} = 28$) and A-level group had 54 participants ($n_{female} = 16, n_{male} = 38$). No participant chose to self-define their gender identity.

2. Group B

All students enrolled in a physics degree (including dual degrees) are required to take the compulsory module "Mechanics, Waves and Optics" in their first semester at the University of Sheffield. This established course is split into four subunits (mechanics, waves, optics, and special relativity), each of which are taught by a different lecturer through didactic lectures (12 lectures per subunit) with additional weekly 1 h small group tutorials. The reduced FCI v95 was delivered via the Blackboard Virtual Learning Environment at three time points:

(1) **Precourse**—during the first week of first semester,
(2) **Postcourse**—the week immediately after completion of the mechanics component of the module (6 weeks after precourse test),
(3) **Delay**—three months after postcourse test.

Ninety-three students out of the class of 174 completed the FCI at all three time points and thus make up group B ($n_{female} = 26, n_{male} = 67$). They used a unique identification number to allow longitudinal analysis to be undertaken but this also prevented the use of engagement measures or course attainment as a covariate within analysis. Students were not provided with their results after any of the tests. Scores did not contribute to any of their university course grades and students were informed that participation was voluntary. As this group comprises of participants with a similar educational background and current educational situation they were only asked to voluntarily provide their gender identity after the delay test. An independent samples Mann-Whitney test showed no statistically significant difference between A-level grades of male and female undergraduate students in group B. The average grades achieved for both male and female groups was one grade above the degree entrance requirement, typically an AAB (compared to the entry requirement of ABB).

III. RESULTS

A. Effect of gender on scores for each level of qualification

A statistically significant difference was found between males and females for GCSE or O level (group A), A level (group A), and precourse (group B), with males outperforming in all cases. For group B this gender difference persisted immediately after completion of a mechanics course as well as 3 months after completion of course. The results of all analyses are shown in Table I. For both male and female groups there is an increase in average FCI score with increasing level of qualification in physics. This is in agreement with the work of Coletta and Phillips who found a correlation between SAT and FCI scores at four universities in the U.S. [43].

Cohen's d [44] was calculated to indicate the size of the effect of gender for all of these comparisons. In all cases the effect size is considered to be large when using Cohen's original suggested interpretation. However, more recent work by Rodriguez and colleagues [45] cautions that effect sizes are not absolute indicators in physics education research. They note that care must be taken when comparing effect sizes observed in different studies published across the research field due to the uniqueness of the specific groups and conditions being compared.

B. Effect of gender on FCI and retention

Wilcoxon signed rank tests between precourse and postcourse scores as well as between precourse and delay scores showed a statistically significant difference for both

TABLE I. Average scores for each group split by gender. Comparisons were conducted using an independent samples Mann-Whitney test and a statistically significant difference was found in all cases. A power analysis indicated that all comparisons were sufficiently powered (GCSE = 0.927, A level = 0.971 and 0.981, 0.976, and 0.923 for precourse, postcourse, and delay, respectively).

| | | Average ± standard deviation (%) | | | | | |
		Female	Male	p	U	Z	Cohen's d
Group A	GCSE/O-level	20.4 ± 15.1	34.2 ± 19.1	0.001	313	−3.437	0.84
	A-level	43.0 ± 19.8	65.8 ± 19.9	0.001	134	−3.234	1.17
Group B	Precourse	68.1 ± 15.8	82.4 ± 15.2	<0.001	423	−3.863	0.94
	Postcourse	80.8 ± 13.5	89.8 ± 8.1	0.002	510	−3.166	0.92
	Delay	80.5 ± 14.2	88.9 ± 9.1	0.007	565	−2.668	0.79

TABLE II. Two measures of the effect of instructions, Cohen's d and normalized learning gain $\langle\langle g \rangle\rangle$, disaggregated by gender. Precourse and postcourse as well as precourse and delay comparisons were statistically significant for both females and males ($p < 0.001$). Postcourse and delay showed no statistically significant difference but are included here for comparison.

	Precourse to postcourse		Precourse to delay		Postcourse to delay	
	Cohen's d	$\langle\langle g \rangle\rangle$	Cohen's d	$\langle\langle g \rangle\rangle$	Cohen's d	$\langle\langle g \rangle\rangle$
Female	0.88	0.40 ± 0.11	0.84	0.39 ± 0.11	0.02	-0.012 ± 0.002
Male	0.61	0.42 ± 0.09	0.51	0.36 ± 0.08	0.11	-0.09 ± 0.01

female and male groups ($p < 0.001$ in all cases, all powered at 0.99). Both learning gains and effect sizes are calculated for the tests with statistical significance. These results are summarized in Table II. Comparison between postcourse and delay scores showed no significant difference for either male or female cohort groups. This result is in line with previous studies [32], however here the retention is also shown to be independent of gender.

Although the normalized learning gains [30] defined by Eq. (1) are included in Table II to allow for comparison with other studies these values should be interpreted with care. Participants in group B demonstrated a high average on the FCI when compared to a majority of studies conducted in the U.S., which is likely due to the specialist nature of UK degrees, namely, that all 93 participants were registered for a full physics degree course. These high averages mean the normalized learning gain is more susceptible to skew and therefore Cohen's d is a better representation of the effect observed.

The effect size between precourse and both postcourse and delay are large for female participants and medium large for male participants. The normalized learning gains suggest that there is little difference between the female and male groups, in contradiction to the statistical analyses above, however as Rodriguez *et al.* [45] demonstrate, the effect size is the more preferred measure compared to normalized learning gain.

IV. DISCUSSION

This study adds to the wide pool of existing research and shows that a gender gap and its subsequent narrowing following instruction also exists within UK-based physics undergraduates. The initial FCI score is considerably higher in this work compared to research undertaken in the U.S., however this is likely due to the differences between the U.S. and UK undergraduate courses, namely, that in the UK students are enrolled on a single degree programme from the start rather than selecting minors and additional credit courses.

This gender gap is also seen at different levels of education in physics from compulsory secondary school education to first year undergraduate students, and the magnitude of the effects are large in all cases. More interestingly, the effect size before and immediately after instruction in mechanics are very similar (0.94 and 0.92, respectively), however the effect decreases slightly after a 3 month delay. This decreased effect size of 0.79 shown in Table I is still statistically significant and suggests the gender gap may decrease with longer time delays however at present no other studies have investigated the gender gap over time and, consequently, the reasons behind this slight decrease are a matter of interest. It may be that different levels of male and female attrition or study participation may introduce a sampling bias or that after three months female students have developed more of a sense of belonging which is known to affect attainment and engagement [46]. Furthermore, there have been many attempts to reduce the gender gap through interactive engagement techniques [37] and the results found here suggest that educators should consider the time between instruction and testing particularly when scheduling formative assessments.

Finally, this work also shows that FCI scores persist after instruction in agreement with the U.S. based study by Francis, Adams, and Noonan [32]. This retention is also shown to be independent of gender with females showing a large effect size gain whereas males only a medium-large gain. This effect is obscured if only the normalized learning gain is considered [30] due to the large initial FCI scores. Despite the larger gains for female participants the average score remains 9% lower than male participants following instruction. It would be interesting to examine whether the use of interactive engagement methods that have been shown to reduce the gender gap remain effective after instruction is complete.

[1] S. Greenfield, J. Peters, N. Lane, T. Rees, and G. Samuels, SET Fair: A report on women in science, engineering and technology, The Greenfield Review, commissioned for the department of Trade and Industry, London (2002).

[2] National Academy of Science, *Rising Above The Gathering Storm: Energizing and Employing America for a Brighter Future* (National Academy Press, Washington, DC, 2005).

[3] S. J. Ceci, D. K. Ginther, S. Jahn, and W. M. Williams, Women in academic science: A changing landscape, Psychol. Sci. Publ. Interest **15,** 75 (2014).

[4] E. Smith, Women into science and engineering? Gendered participation in higher education STEM subjects, Br. Educ. Res. J. **37,** 993 (2011).

[5] D. N. Beede, T. A. Julian, D. Langdon, G. McKittrick, B. Khan, and M. E. Doms, *Women in STEM: A Gender Gap to Innovation*, Issue Brief No. 04–11 (Economics and Statistics Administration, Washington, 2011).

[6] E. Smith and S. Gorard, Is there a shortage of scientists? A re-analysis of supply for the UK, Brit. J. Educ. Res. **59,** 159 (2011).

[7] B. Auyeung, M. V. Lombardo, and S. Baron-Cohen, Prenatal and postnatal hormone effects on the human brain and cognition, Pflügers Archiv-Euro. J. Physiol. **465,** 557 (2013).

[8] E. Seymour and N. M. Hewitt, *Talking about Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO, 1997).

[9] D. F. Halpern, C. P. Benbow, D. Geary, R. Gur, J. Hyde, and M. A. Gernsbacher, The science of sex differences in science and mathematics, Psychol. Sci. Publ. Interest **8,** 1 (2007).

[10] P. Labudde, W. Herzog, M. P. Neuenschwander, E. Violo, and C. Gerber, Girls and physics: teaching and learning strategies tested by classroom interventions in grade 11, Int. J. Sci. Educ. **22,** 143 (2000).

[11] J. Taylor, Sexist bias in physics textbooks, Phys. Educ. **14,** 277 (1979).

[12] A. Kitchenham, Vive la difference: Gender, motivation, and achievement, School Libraries in Canada **22,** 34 (2002).

[13] J. B. Kahle and J. Meece, Research on gender issues in the classroom, in *Handbook of Research on Science Teaching and Learning* (Macmillan, New York, 1994), pp. 542–557.

[14] A. Zohar and D. Sela, Her physics, his physics: Gender issues in Israeli advanced placement physics classes, Int. J. Sci. Educ. **25,** 245 (2003).

[15] P. Laws, P. Rosborough, and F. Poodry, Women's responses to an activity-based introductory physics program, Am. J. Phys. **67,** S32 (1999).

[16] M. Schneider, Encouragement of women physics majors at Grinnell College: A case study, Phys. Teach. **39,** 280 (2001).

[17] E. Bagno and B.-S. Eylon, From problem solving to a knowledge structure: An example from the domain of electromagnetism, Am. J. Phys. **65,** 726 (1997).

[18] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics Deutsch, Eur. J. Sci. Educ. **1,** 191 (1979).

[19] J. H. Larkin, Cognition of learning physics, Am. J. Phys. **49,** 534 (1981).

[20] L. N. Walsh, R. G. Howard, and B. Bowe, Phenomenographic study of students' problem solving approaches in physics, Phys. Rev. ST Phys. Educ. Res. **3,** 020108 (2007).

[21] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, Phys. Rev. ST Phys. Educ. Res. **3,** 020101 (2007).

[22] T. Byun and G. Lee, Why students still can't solve physics problems after solving over 2000 problems, Am. J. Phys. **82,** 906 (2014).

[23] E. Kim and S.-J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70,** 759 (2002).

[24] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30,** 141 (1992).

[25] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, Phys. Rev. ST Phys. Educ. Res. **11,** 020106 (2015).

[26] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69,** S12 (2001).

[27] W. J. Leonard, R. J. Dufresne, and J. P. Mestre, Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems, Am. J. Phys. **64,** 1495 (1996).

[28] R. K. Thornton and D. R. 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 (1998).

[29] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30,** 159 (1992).

[30] 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).

[31] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. J. Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80,** 638 (2012).

[32] G. E. Francis, J. P. Adams, and E. J. Noonan, Do they stay fixed? Phys. Teach. **36,** 488 (1998).

[33] D. Huffman and P. Heller, What does the force concept inventory actually measure? Phys. Teach. **33,** 138 (1995).

[34] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33,** 503 (1995).

[35] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8,** 020105 (2012).

[36] A. Savinainen and P. Scott, Using the Force Concept Inventory to monitor student learning and to plan teaching, Phys. Educ. **37,** 53 (2002).

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

[38] 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 (2007).

[39] J. L. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).

[40] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).

[41] G. L. Erikson and L. J. Erickson, Females and science achievement: Evidence, explanations, and implications, Sci. Educ. **68**, 63 (1984).

[42] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

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

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

[45] I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, Phys. Rev. ST Phys. Educ. Res. **8**, 020103 (2012).

[46] K. Rainey, M. Dancy, R. Mickelson, E. Stearns, and S. Moller, Race and gender differences in how sense of belonging influences decisions to major in STEM, Int. J. STEM Educ. **5**, 10 (2018).