Uniped
Tidsskrift for universitets- og høgskolepedagogikk
RESEARCH PUBLICATION

UNIVERSITETSFORLAGET

Årgang 43, nr. 1-2020, s. 19–32
ISSN Online: 1893-8981
https://doi.org/10.18261/issn.1893-8981-2020-01-03

Gender Differences in Norwegian Engineering Students' Understanding of Newtonian Mechanics

Jonas R. Persson
Associate professor, Department of Physics, Norwegian University of Science and Technology
jonas.persson@ntnu.no

Abstract
An investigation of the gender gap in engineering students' understanding of Newtonian physics has been done. The results indicate a noticeable gender gap, in addition to an effect due to differences in preparation, i.e. physics courses in secondary school. The gender gap after university instruction decreases in the case of a higher degree of preparation, while the gender gap is unchanged or widens for students with a lower degree of preparation after university instruction.

Keywords
Concept inventory, Newtonian mechanics, gender gap, physics

Introduction
There are concerns about female participation in physics and technology education, as there is still a majority of men who choose an education in physics and technology in Norway, even if the number of women is increasing. In addition, the performance of female students has also been a subject of discussion as well as a concern. Here we examine the gender differences in performance in Newtonian mechanics within introductory physics courses at the Norwegian University of Science and Technology (NTNU), using the Force Concept Inventory (FCI). Developed by Hestenes and co-workers at Arizona State in the early 1990s, the FCI has become the gold standard diagnostic test of conceptual understanding in physics (Hestenes et al. 1992). Its value was shown in Hake's landmark paper in 1998 (Hake, 1998), where the instrument was used to show the efficiency of different instructional methods using a sample of nearly 6000 students, making it widely accepted and used in physics education research.

A consensus has been developed regarding the existence of a performance gender gap in the FCI. The performance gender gap has previously been investigated by several studies at US institutions and in the UK, using the FCI instrument in introductory physics courses. The standard methodology is to assess students with the instrument prior to and then after the instruction, in the so-called pre-post methodology. Lorenzo et al. (2006) presented extensive data from Harvard students undertaking an introductory calculus-based course between 1990 and 1997. They found a statistically significant gender gap in the pre-tests, where the mean score for male students was consistently better than the mean score for

female students. In addition, they found that certain instructional methodologies were more effective at reducing (or even eliminating) the gap. Especially effective was the interactive engagement style of instruction with peer instruction and discussions, developed by Eric Mazur in the late 1990s, as a response to students' initial performance on the FCI at Harvard.

Similar studies at the University of Colorado, Pollock et al. (2007) and Kost-Smith et al. (2009) found that it was not sufficient to use engagement instructional methodologies to close or eliminate the gap in FCI performance. There are, however, examples where the gap widens despite an overall improvement in both pre- and post-test results. This suggests that other factors, such as gender differences in preparation and background, could be important. A large study by Docktor and Heller (2008), with results from 40 separate classes, 5500 students and 22 different instructors, confirm a significant gender gap in pre-tests, which persists in post-tests. However, the gap changes in a range from +7% (gap widens) to −6% (gap narrows), in individual classes.

Madsen et al. (2013) have discussed the question of the gender gap of the conceptual inventories in physics and the factors contributing to it. In the case of FCI, Madsen et al. (2013) found that sociocultural factors[1] had an impact on the gender gap. It has been suggested that the test items may be written in a way that favours men. (McCullough and Meltzer, 2001). Alternative item context, more "feminine" and everyday wording showed differences on individual items, but not the overall result. This motivated Traxler et al. (2017) to investigate the gender fairness of FCI, based on item response theory (IRT) and treating gender as a factor. Their results point to a number of items as being "problematic", which should be removed in the analysis. Their results indicate that the removal of these "unfair" items will lessen the "gender gap" as items with gender bias are removed. However, the result will probably depend on the test sample, which is why it may not be applicable to different test samples.

Even if FCI is well studied, there remain questions on the structure of the test. The persistent gender is one that is not fully understood. Is the test itself intrinsically biased, in wording, examples or format? The research of Dawkins et al. (2017) suggests that the structure and scaffolding of the test might be important.

Even if FCI has been extensively studied, the origin of the gender gap is not fully understood as indicated by previous studies. We are motivated to examine a possible gender gap in the Norwegian context, in part due to the lack of consensus in the literature, but also due to the differences in the style of university education and the preparation prior to university studies. We apply our investigation to a cohort of students in different Master of Technology studies, in which female students are under-represented.

The Norwegian upper-secondary school system allows for a choice on the level of physics, making it possible to start university studies in STEM with only the first of two possible physics courses in upper secondary school. This is expected to give the students who start their university studies with both courses, an advantage in the performance on FCI and in their studies. This will give an indication of how important the level of preparation is for a possible gender gap, something that has not been addressed in previous studies at the university level. That is, will the gender gap decrease with a higher degree of preparation in secondary school?

Our study has the following aims:

a. To evaluate if there is an initial performance gender gap in introductory physics courses at a Norwegian university.

1. For example: level of endorsement of gender stereotypes, students' beliefs in their answers.

b. To evaluate if there exists an initial performance gap due to the background, i.e. level of physics in upper secondary school.
c. To evaluate the persistence of the possible performance gaps after instruction.
d. To evaluate differences in the performance of test items, due to gender and background.

This paper is organised as follows. The next section describes the students and study-programs. The methodology to obtain and analyse test data is presented in Section 3. The results are presented in Section 4 and in the final section, a discussion of the results and implications are presented.

Methodology

FCI

The Force Concept Inventory (FCI) instrument (Hestenes et al. 1992) is a multiple-choice test, designed to assess student understanding of the most basic concepts in Newtonian physics. The test has 30 questions covering different areas of understanding: Kinematics, Newton's Laws, the superposition principle, and types of forces. A sample question is given in Figure 1. Each question has only one correct Newtonian answer, with four distractors based on known student's misconceptions. A low score indicates that the student has an Aristotelian understanding while a high score (typically, above 60% correct) indicates a Newtonian understanding. Angell and collaborators at the University of Oslo (Angell 2012) translated the Norwegian version of FCI used in the study.

A stone dropped from the roof of a single story building to the surface of the earth:

(A) reaches a maximum speed quite soon after release and then falls at a constant speed thereafter.

(B) speeds up as it falls because the gravitational attraction gets considerably stronger as the stone gets closer to the earth.

(C) speeds up because of an almost constant force of gravity acting upon it.

(D) falls because of the natural tendency of all objects to rest on the surface of the earth.

(E) falls because of the combined effects of the force of gravity pushing it downward and the force of the air pushing it downward.

Figure 1. Sample question from the FCI.

Physics in the Norwegian education context

Physics is part of an integrated school science subject in compulsory school (Years 1–10) and in the first year of upper secondary school. The students in the general pre-academic specialisation can take physics and other science subjects during the last two years. In general, Physics 1 and/or another science subject (biology and/or chemistry) is required for admission to science studies at the universities. For Medical School and Master in Technology programmes, Physics 1 is required. This can explain a relatively large enrolment in Physics 1, where students want to keep as many options for future studies as possible open. The advanced Physics 2 course is recommended for university studies in science and technology, but the enrolment is about 30–50% lower than Physics 1. The system thus allows students to start science or technology studies with different knowledge bases in physics. However, a majority (~80%) of the students in this study have had Physics 2 in upper secondary school.

University context and courses

The test was given in three different courses with student groups, both as a pre-(instruction) and post-(instruction) test 2012–2015. The courses were all traditional calculus-based introductory physics courses. As all engineering students at NTNU must take at least one course in physics, it was possible to administer FCI to both physics masters and non-physics masters. However, different physics courses are given to different master's programs, but all courses contain approximately the same amount of content relevant for the FCI during lectures and all use the same textbook. There were no major differences in the way the relevant material was presented in the lectures.[2] The main difference was in the time spent in lecture and degree of mathematical sophistication. The basic concepts, which FCI is designed to investigate, do not differ in the courses.The three groups consisted of students in different physics courses: Mechanical Physics (TFY4145) for Physics Masters; Physics (TFY4104) for Master students in Marine Technology, Industrial Economics and Technology Management, and Mechanical Engineering; and Physics (TFY4115) for Master students in Electronics, Engineering Cybernetics and Nanotechnology. It should be noted that TFY4104 and TFY4115 also include electromagnetics and thermodynamics, respectively.

Implementation of FCI

The FCI pre-test was administered during the first week of the physics courses, with a large attendance giving a large number of answers. The post-test was administered during the last three weeks of the semester, where attendance dropped as well as the number of answers. The FCI was answered anonymously, where the students were given an identification number to keep for the post-test. About 50% of the students who did the post-test used their identification number. The total number of participants in different cohorts (Male, Female, Physics 1 and Physics 2) are given in Table 1. In addition to results for all answers, we also use a group of "matched" answers, that is the students that we could identify who took both pre- and post-tests.

The FCI was administered in courses from 2012 to 2015, in order to obtain a large number of answers.

Table 1. Student cohorts and the total number of participants in the FCI test.

Cohort	Number of participants	
	Pre-test	Post-test
Physics 1		
Male	132	29
Female	166	39
Physics 2		
Male	870	240
Female	332	81
Total	**1539**	**402**

2. As indicated by the lecturers' notes.

Results

Quantitative analysis

The FCI was administered in each course before and after relevant instruction. All the answers were divided into different cohorts based on gender and courses in upper secondary school (Table 1).

The results for the whole cohorts (all answers) and the mean performance are presented in Table 2. For the analysis of the male and female sub-cohorts, we define the gender gap as the difference between the male and female mean scores. The convention used is that positive gaps imply male students doing better than female and vice versa.

For students undertaking both the pre- and post-test, it is possible to calculate a cohort-averaged normalised gain <g> defined as

$$\langle g \rangle = \frac{\langle x \rangle_{post} - \langle x \rangle_{pre}}{100 - \langle x \rangle_{pre}}$$

This normalised gain is often considered as a measure of the effectiveness of instruction, representing the fractional improvement in understanding (Hake, 1998). The cohort of students identified as taking both pre- and post-test (matched group) is presented in Table 3.

Table 2. Student cohorts and performance on the FCI for all participants. Values in parentheses are the standard error of the mean.

Cohort	Number of participants	Performance (%)		
		Pre-test	Number of participants	Post-test
Physics 1				
Male	132	61.5(1.7)	29	75.0(3.8)
Female	166	44.9(1.4)	39	56.0(3.3)
Gender gap		16.6		19.0
Cohen's delta		0.89		0.92
Physics 2				
Male	870	80.1(0.7)	240	85.1(1.0)
Female	332	62.6(1.2)	81	74.9(2.1)
Gender gap		17.5		10.2
Cohen's delta		0.83		0.61

Table 3. Student cohorts and performance on the FCI for individuals taking both pre- and post-test (matched group). Values in parentheses are the standard error of the mean.

Cohort	Number	Performance (%)		<g>
		Pre-test	Post-test	
Physics 1				
Male	14	67.9(4.9)	75.7(5.8)	0.24
Female	18	51.9(4.1)	62.2(4.9)	0.21
Gender gap		16.0	13.5	
Cohen's delta		0.90	0.64	
Physics 2				
Male	122	80.2(1.5)	85.0(1.4)	0.24
Female	53	70.1(3.0)	76.6(2.8)	0.22
Gender gap		10.1	8.4	
Cohen's delta		0.55	0.49	

From the result in Tables 2 and 3 is it evident that there is a significant difference between the different cohorts in both the pre- and post-tests. The cohorts with only the first physics course in upper secondary school score significantly lower. There also exists a significant gender gap in all cohorts with medium to large effect sizes (Cohen's delta).

When a comparison is made for all participants (Table 2), the gender gap narrows (from 17.5% to 10.2%) for the group with both physics courses in upper secondary school. However, one should be aware that the scores are approaching the effective ceiling value for this group. A FCI score of 85% has been suggested as the Newtonian Mastery threshold (Hestenes & Halloun 1995), suggesting that the participants reached high levels of understanding. It is noteworthy that the male students in this group are close to this prior to instruction, making additional improvement difficult. The Newtonian thinking has passed the entry threshold (60%) (Hestenes & Halloun 1995) for both male and female students with Physics 2.

In the group with only the first physics courses in upper secondary school, we see a similar pattern, the male students outperforming the female students, in this case, the gender gap seems to widen for the whole cohort (all answers). However, the number of participants is lower, which is why this might influence the result. Improvement in gain is similar to the other group. The degree of Newtonian thinking is low among female students, with the mean not reaching the entry threshold (60 %) (Hestenes & Halloun 1995), either in the pre- and post-tests. Even if it is possible to obtain interesting results using the whole group, the limited number of students and the fact that not all did the post-tests pose some questions that need to be considered about the validity of the result.

The tests were administered during lectures, and this will evidently cause those students who do not attend lectures to be excluded in the post-test. The general feeling is that the students not attending lectures are predominantly those struggling and this will mean that the pre-test data is for the whole cohort, while the post-test will probably probe the high achieving cohort. This can also be inferred by the slightly higher result in the pre-test for the

matched pre-post group (Table 3), something that seems to be more evident for female students. That is, the weaker students (low FCI score in the pre-test) will not have answered the post-test. The pre-test results will, therefore, probe all students' knowledge prior to instruction; while to investigate effects due to instruction, only students doing both pre- and post-test will be investigated. The loss of students with low pre-test FCI scores is unfortunate, as information on those students is needed to design proper reformed instruction.

In Table 3, only students participating in both pre- and post-tests are included. The effect of excluding students does not seem to influence the results for male Physics 2 students in the pre-test and both genders in the post-test. For these groups, the gender gap narrows as well. The pre-test for female students is significantly improved, which supports the hypothesis that "weaker" students skip lectures towards the end of the course. However, this can also be interpreted as a sign of gender-biased instruction.

The limited amount of data for the Physics 1 groups can only be used for a general indication of trends. We see an improvement of pre-test scores, explained by the filtering of weak students. As with the Physics 2 group the gender gap narrows. The level of Newtonian thinking reaches the entry level for females in the post-test but is still quite low.

The problem of not reaching a sensible level of Newtonian thinking is quite alarming as this might affect the students' future understanding of other subjects in their studies. The clear difference of results indicates that the problem might be based on gender issues.

From the results on the gain in different cohorts, one finds that the gain for the instruction at NTNU is as expected of traditional instruction (Hake, 1998). This gives an indication that the instruction on the subject addressed in FCI should be changed, to better address the students' difficulties. This can also motivate to introduce reformed instruction in the courses.

Results from unbiased FCI

Traxler et al. (2017) found in their analysis that the FCI might not be gender fair, and recommended that 10 items should be removed. However, their analysis is based on three samples where the effect of the "unbiased" FCI differed with a significant effect in one sample. This raises the question of whether this effect is context-dependent and therefore interesting to investigate in another educational context. Removing items (6, 9, 12, 14, 15, 21, 22, 23, 24, 27) produces a 20-item instrument, which might decrease the validity of the instrument, but in this case, we are interested in the effect of removing gender-biased items.

Table 4. Student cohorts and performance on the "unbiased" FCI for all participants. Values in parentheses are the standard error of the mean.

Cohort	Number of participants	Performance (%)		
		Pre-test	Number of participants	Post-test
Physics 1				
Male	132	57.8(1.8)	29	77.6(4.2)
Female	166	46.1(1.5)	39	57.1(3.6)
Gender gap		11.7		20.5
Cohen's delta		0.61		0.92

Cohort	Number of participants	Performance (%)		
		Pre-test	Number of participants	Post-test
Physics 2				
Male	870	78.9(0.6)	240	84.8(1.0)
Female	332	64.1(1.2)	81	77.3(2.2)
Gender gap		14.8		7.5
Cohen's delta		0.82		0.42

The results for the "unbiased" FCI for all students is given in Table 4. The result is inconclusive as to the gaps in the different cohorts' decreases in three cases and increases in one (Physics 1, Post-test) compared to "biased" FCI (Table 2). The values in the Physics 2 pre-test and the Physics 1 post-test are almost the same when comparing the effect (Cohen's delta) with Table 2, thus indicating that "unbiased" FCI have no effect on the gender gap. However, the decrease is clear in other cases. This poses a problem, as there might be a number of reasons for this and uncertainty in how to find the real reason. One way is to compare with the students who answered both pre- and post-tests.

Table 5. Student cohorts and performance on the "unbiased" FCI for individuals taking both pre- and post-test(matched group). Values in parentheses are the standard error of the mean.

Cohort	Number	Performance (%)		<g>
		Pre-test	Post-test	
Physics 1				
Male	14	66.1(4.8)	74.0(6.3)	0.23
Female	18	50.3(4.8)	62.8(5.5)	0.34
Gender gap		15.8	11.2	
Cohen's delta		0.82	0.48	
Physics 2				
Male	122	79.9(1.5)	84.9(1.4)	0.25
Female	53	71.5(3.3)	79.1(2.9)	0.27
Gender gap		8.4	5.8	
Cohen's delta		0.45	0.31	

Comparing the results for the pre-post-test students in Table 5 with the result in Table 4 shows clear effects in all cohorts. Note the high gain for Physics 1 females, which is due to the removal of certain items in the "unbiased" FCI. The decrease in gender gap compared with FCI in Table 3 indicates that there might be a gender bias in FCI in this context. There might also be an underlying connection between preparation and gender, with origin from upper secondary school or earlier. However, to investigate this in detail would be outside the scope of this article and is left for further research.

Item analysis

Observing the existence of a gender gap opens the question of whether the gap is due to an outperformance across the entire test or if there are certain questions where the male students outperform the female students. In this analysis, only students who answered both pre- and post-tests are included.

By plotting the fraction of male and female students getting an item correct, it is easy to find the problematic items. In Figures 2–5, the fraction of correct answers is given for different groups and test. As can be seen, the data confirms that a larger fraction of the male students get most items correct compared with the female students.

There are items that tend to have a rather large gender gap. Comparing the results for individual items we observe that items 21, 22 and 23 form a cluster in addition to item 14, with fewer correct answers for women compared to men in all cases. It should be noted that items 14 and 23 are among those with the largest gender gap as reported by Docktor et al. (2008) and Madsen et al. (2013). These items, together with items 6, 8, 12, 14 and 21, describe motion in two dimensions. Items 21 and 23 are similar to item 8 but deal with a continuous force, where item 8 describes a momentary force. The gender gap for item 8 decreases substantially for the Physics 2 cohort, while the decrease for the other items is smaller. For the Physics 1 cohort, the gap is more stable. Item 12, motion in two dimensions, is treated extensively in Physics 2, which explains why it is a problematic item in the Physics 1 cohort, due to lack of pre-training. However, the result indicates that a multidimensional context may be a partial explanation for some of the gender gap as shown by Dawkins et al. (2017). It is also noteworthy that item 6, which Traxler et al. (2017) removed, generally has a relatively small gender gap (<10%) in this context.

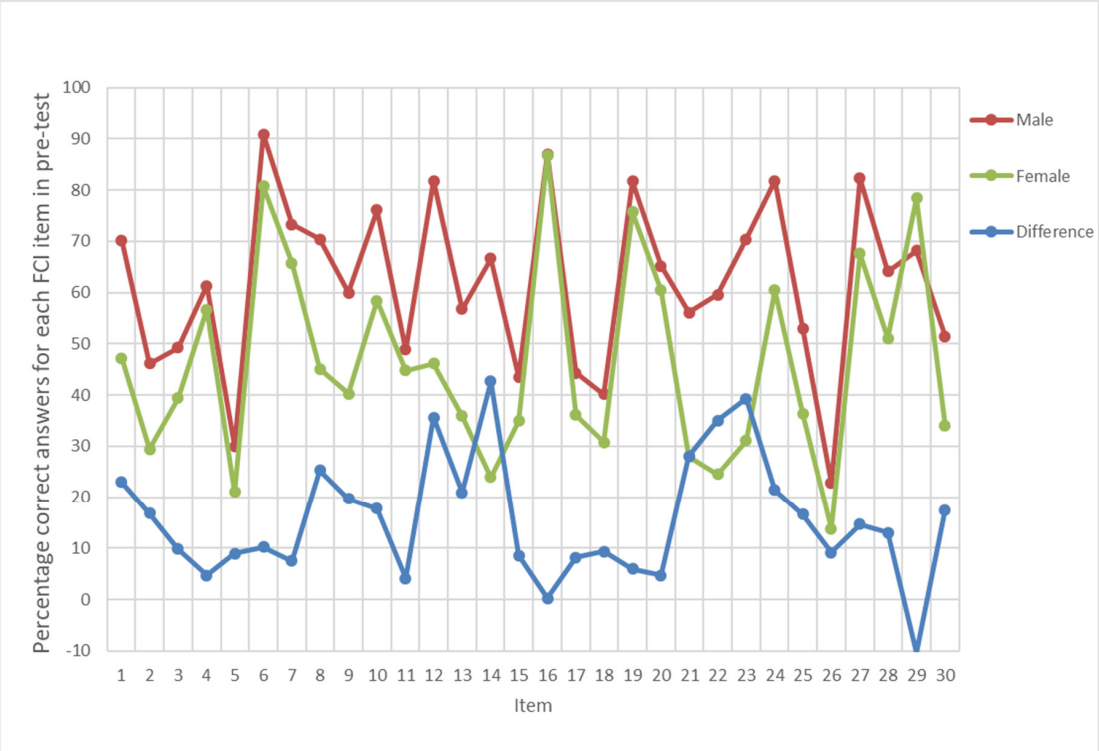

Figure 2. Percentage of correct answers on individual items for male and female students in the Physics 1 pre-test. The difference between male and female students is also given.

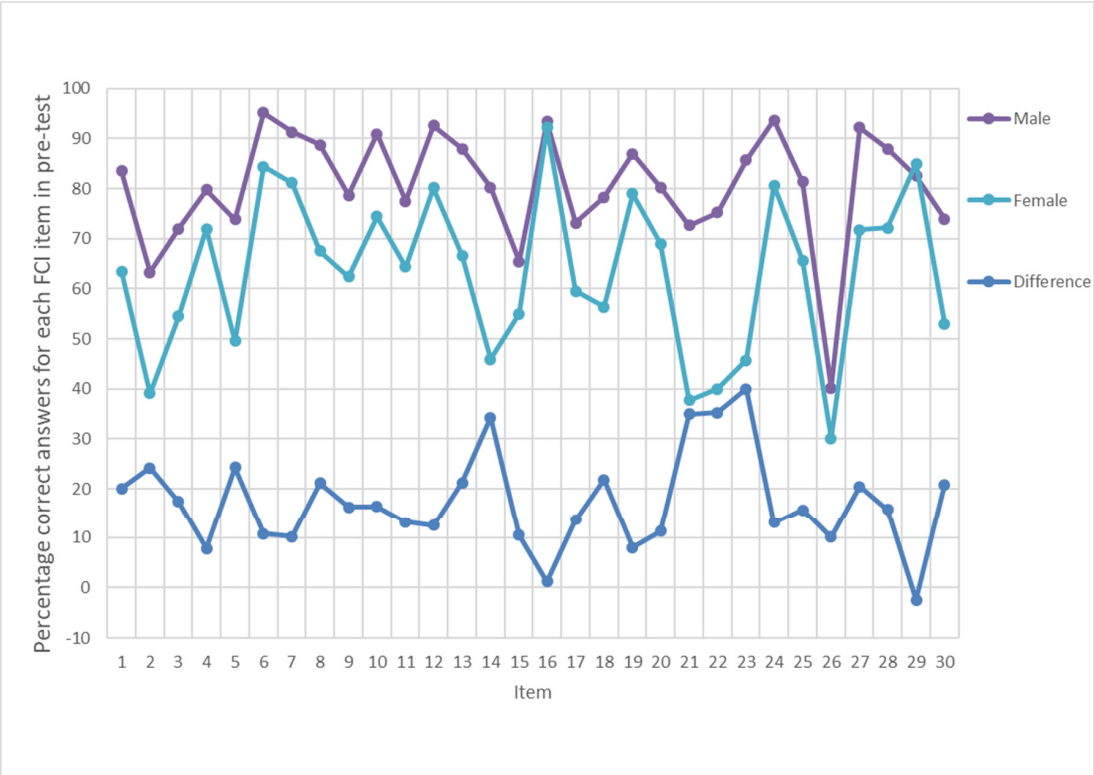

Figure 3. Percentage of correct answers on individual items for male and female students in the Physics 2 pre-test. The difference between male and female students is also given.

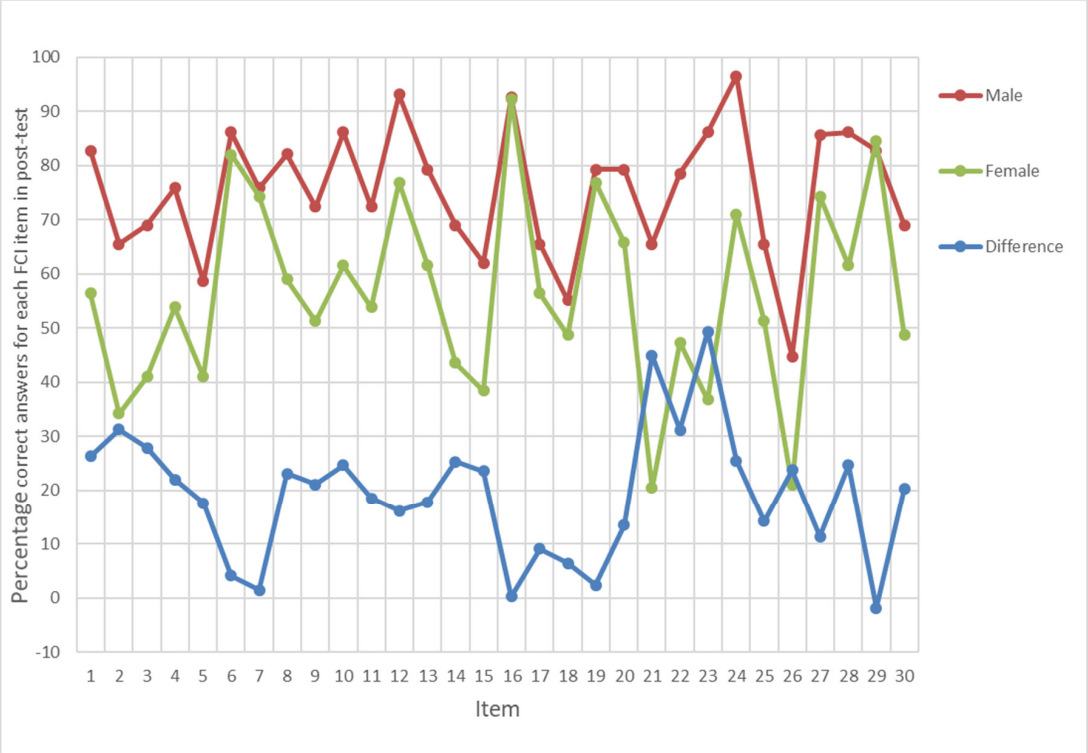

Figure 4. Percentage of correct answers on individual items for male and female students in the Physics 1 post-test. The difference between male and female students is also given.

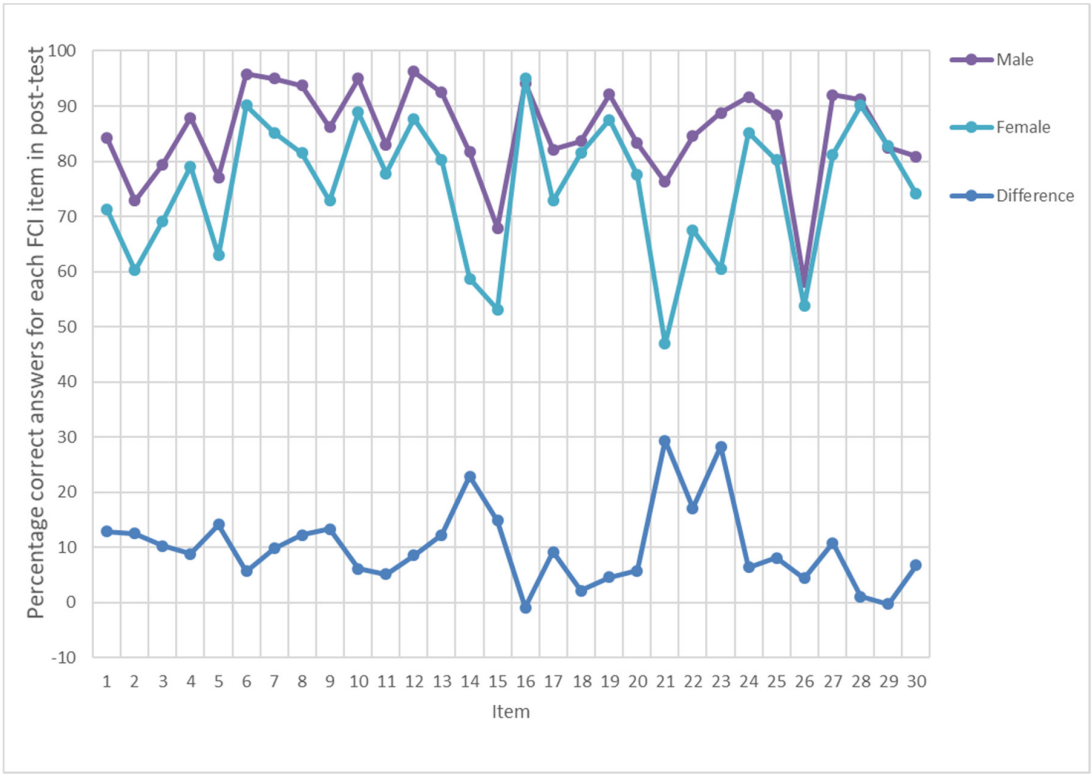

Figure 5. Percentage of correct answers on individual items for male and female students in the Physics 2 post-test. The difference between male and female students is also given.

The gender gap can be investigated further by studying groups of items. The individual items may give an indication of both the specific conceptual understanding and the semantic and experience-based understanding of the item. In this case, the conceptual understanding is more important and the items are grouped in distinctive conceptual dimensions. The original conceptual dimensions of Hestenes et al. (1992), place items in multiple categories, but in our study, we require items to be placed in only one category. We, therefore, use five conceptual categories: Newton's first law, Newton's second law, Newton's third law, Kinematics and Identification of force. The first four are identical to Hestenes' dimensions, while Identification of force contains items from Hestenes' Kind of Forces dimension.

The gender gaps in the different categories and groups are given in Table 6. The overall pattern of an existing gender gap is clear, even when considering the "unbiased" FCI (Traxler, et al. 2017).

The increasing gap for the Physics 1 group is not homogeneous in all categories; while some gaps seem to be stable, others increase. It seems like Newton's third law is especially difficult, with an increasing gap of about 10 percentage points. The effect of the "unbiased" FCI in the category "Newton's 2nd law" for Physics 1 seems spurious and should be investigated further. In the Physics 2 group, the gender gap decreases in all.

Table 6. Gender gap in different categories for different groups.

Category	Items in FCI	Physics 1 Pre	Physics 1 Post	Physics 2 Pre	Physics 2 Post
Newton's 1st law	6,7,8,10,12,21,23	23.4	23.4	20.9	14.3
"unbiased"	7,8,10	16.9	16.4	15.9	9.4
Newton's 2nd law	3,9,22,24,25,26,27	18.1	22.2	18.3	10.1
"unbiased"	3,25,26	11.8	22.0	14.4	7.6
Newton's 3rd law	4,15,16,28	6.7	17.6	8.9	6.0
"unbiased"	4,16,28	6.0	15.7	8.3	3.0
Kinematics	1,2,14,19,20	18.6	19.8	19.6	11.7
"unbiased"	1,2,19,20	12.6	18.4	15.9	9.0
Identification of Force	5,11,13,17,18,29,30	8.4	10.0	16.1	11.6

The increasing gender gap for the Physics 1 group seems to indicate that previous preparation is important, at least when it comes to certain aspects of forces. As this is not observed in the Physics 2 group, the gender gap probably depends on a combination of different causes, such as preparation, gender-biased teaching[3] and the test itself.

Discussion

The results present a picture of a noticeable gender gap in both pre- and post-instruction FCI results. Even when considering gender bias effects and excluding items to increase gender fairness (Traxler et al., 2017), the gap is still noticeable but slightly decreasing. An effect due to prior preparation in secondary school is clearly visible. The trend found in the "matched group" (Table 5) indicates that preparation is an important factor when attempting to decrease the gender gap. However, the results do not give any answers on what to improve in the preparation.

It is interesting to note the large difference in item 12 between Physics 1 and 2, which can be explained by the fact that similar pictures are shown in the Physics 2 textbooks. Showing that familiarity might be important. One can speculate that time spent on difficult concepts, such as force, could be an important factor. A superficial treatment without a focus on understanding or "hands-on" activities, especially for individuals without previous tactile experience,[4] might be an explanation.

However, an analysis of individual items indicates an additional underlying problem. As most of the problematic items are based on motion in two dimensions, gender-based differences in spatial visualisation might be another explanation. That there exists an advantage for males in spatial skills is well documented (Linn and Petersen 1985; Voyer et al. 1995; Halpern 2000; Voyer et al. 2017). Hake (2002) investigated the correlation between gain and spatial score (from the Purdue Spatial Visualisation test [of Rotations] [PSVT:R] [Guay 1977]) and found a small correlation (r=0.24). A difference in both spatial and FCI (both pre- and post) was also noticeable. Similar relationships have also been found by Mac

3. That is, sociocultural aspects in the teaching practice, with a negative bias towards women. For example in wording, choice of examples, et cetera.
4. For example with different toys or other activities during childhood.

Raighne et al. (2015) using the Force Motion Concept Evaluation (FMCE) and PSVT:R. Even if Hake (2002) did not look at individual items and Mac Raighne et al. (2015) used another inventory, it seems likely that spatial skills might explain part of the gender gap in FCI and beyond. As it is a possible cause, more research is needed as well as investigating to what degree training spatial abilities are addressed in schools and university. Even though "feminine" and everyday wording (McCullough and Meltzer 2001) did not affect the gender gap does not mean that the structure of the test is biased. Dawkins et al. (2017) have shown that a higher degree of scaffolding decreases the gender gap, although they are not able to sufficiently explain why the effect of a higher degree of scaffolding in the items should be investigated.

Investigations using the CLASS (Colorado Learning Attitudes about Science Survey) on Physics Majors has shown a difference in attitudes between male and female students (Persson, 2016, 2017). This might indicate a possible underlying connection to the gender gap, where female students had lower self-confidence and belief in their problem-solving ability.

Our results can be used as an argument that Physics 1 is not enough to be admitted to engineering studies, given the relatively low score for this group in the pre-test. The results in the post-test do not fully support such an argument as the students, on average, reach an acceptable level. In addition, the FCI focuses on a small part of the curriculum where specific difficulties have been found. It should, however, be noted that if important skills necessary for future careers in STEM are missing in Physics 1, as might be inferred from these results, a revision of the curriculum is a better way to proceed than to change how students are admitted to engineering studies.

The results indicate that teaching at university decreases the gender gap but does not give any indication on what to change. The study done is solely based on traditional teaching, lectures, exercises and laboratory activities. If familiarity and experience, as speculated, is an important factor, concrete examples with demonstrations during lectures and "hands-on" examples, where the tactile dimension is important, during exercises could be a relatively easy way to address some causes of the gender gap.

The "unbiased" FCI should be investigated further in the Norwegian context, using Classical Test Theory, Item Response Theory and Differential Item Functioning, in order to find gender asymmetries that might be different compared with the US context (Traxler et al., 2017). Studies of the spatial visualisation abilities complementary to FCI and CLASS might also be advantageous. In addition, the study should be repeated with different reformed teaching in order to find what works.

References

Angell, C. (2012). Private communication.

Dawkins, H., Hedgeland, H., & Jordan, S. (2017). Impact of scaffolding and question structure on the gender gap. *Physical Review Physics Education Research, 13*(2), 020117. https://doi.org/10.1103/physrevphyseducres.13.020117

Docktor, J., & Heller, K. (2008). Gender differences in both force concept inventory and introductory physics performance. *AIP Conf. Proc. 1064,* 15–18. https://doi.org/10.1063/1.3021243

Glasser, H. M., & Smith, J. P. (2008). On the vague meaning of 'gender' in education research: the problem, its sources, and recommendations for practice. *Educ. Researcher, 37*, 343. https://doi.org/10.3102/0013189x08323718

Guay, R. B. (1977). *Purdue spatial visualization test-visualization of rotations*. W. Lafayette, IN. Purdue Research Foundation.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64. https://doi.org/10.1119/1.18809

Hake, R. R. (2002). Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization. *Physics Education Research Conference, 8,* 1–14.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. New York: Psychology Press.

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory. *The Physics Teacher, 33,* 502. https://doi.org/10.1119/1.2344278

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30,* 141. https://doi.org/10.1119/1.2343497

Kost-Smith, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physics Review Spec. Top. Physics Educcation Researcg, 5,* 010101. https://doi.org/10.1103/physrevstper.5.010101

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 1479–1498. https://doi.org/10.2307/1130467

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics, 74*, 118. https://doi.org/10.1119/1.2162549

Mac Raighne, A., Behan, A., Duffy, G., Farrell, S., Harding, R., Howard, R., Nevin, E., & Bowe, B. (2015). *Examining the relationship between physics students' spatial skills and conceptual understanding of Newtonian mechanics*. In 6th Research in Engineering Education Symposium: Translating Research into Practice, REES 2015.

Madsen, A., McKagan, S.B., & Sayre, E. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influences the gap? *Physics Review Spec. Top. Physics Education Research, 9,* 020121. https://doi.org/10.1103/physrevstper.9.020121

McCullough, L., & Meltzer, D. E. (2001). Differences in male/female response patterns on alternative-format versions of FCI items. In Cummings K., Franklin S. & Marx J. (Eds.), *Physics Educational Research Conference Proceedings* (pp. 103–106). New York: AIP publishing.

Persson, J. R. (2016). Ändringar i attityder och föreställningar hos första års-studenter i civilingenjörsutbildningen i fysik och matematik vid NTNU. *Uniped, 39*(01), 37–46. https://doi.org/10.18261/issn.1893-8981-2016-01-04

Persson, J. R. (2017). A tale of two cities. *Uniped, 40*(4), 346–360.

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physics Review Spec. Top. Physics Education Research, 3,* 010107. https://doi.org/10.1103/physrevstper.3.010107

Traxler, A., Henderson, R., Stewart, J., Stewart, G., Papak, A., & Lindell, R. (2017). *Gender fairness within the Force Concept Inventory,* ArXiV 1709.00437.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250. https://doi.org/10.1037/0033-2909.117.2.250

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review, 24*(2), 307–334. https://doi.org/10.3758/s13423-016-1085-7