Revisiting the Confidence Gap in University-Level Programming Courses

Leo Leppänen
University of Helsinki
Helsinki, Finland
leo.leppanen@helsinki.fi

Nea Pirttinen
University of Helsinki
Helsinki, Finland
nea.pirttinen@helsinki.fi

Erkki Kaila
University of Helsinki
Helsinki, Finland
University of Turku
Turku, Finland
erkki.kaila@helsinki.fi

Abstract

The confidence gap, that is, the difference in self-assessed confidence between genders regarding the ability to succeed in a given task, has been previously reported and studied in the context of computer science education. In this paper, we investigate the prevalence of the confidence gap during subsequent introductory and advanced Python programming courses – attended by a range of students from CS majors to non-university students taking the courses as free MOOCs – from a variety of perspectives, including student perceptions of what constitutes as 'good' or a 'bad' grade, what they predict as their own grade at the start of the course, and previous exposure to programming in various contexts. Our results provide both an updated snapshot into the evolving confidence gap, and additional data about students' beliefs on how well they are likely to perform in university-level programming courses targeted at first-year students.

CCS Concepts

• **Social and professional topics → Computing education**.

Keywords

confidence gap; self-efficacy; gender; introductory programming; CS1; MOOCs; higher education

ACM Reference Format:
Leo Leppänen, Nea Pirttinen, and Erkki Kaila. 2025. Revisiting the Confidence Gap in University-Level Programming Courses. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2025), June 27-July 2, 2025, Nijmegen, Netherlands.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3724363.3729103

1 Introduction

The existence of the 'gender gap' in computing is unlikely to come as a surprise to anyone even broadly involved with computing education. Much has been written about women's lower participation in computer science, to the point where some authors have identified persistent – even harmful – narratives and frames through which the issue has traditionally been investigated [32]. Often, the gap is attributed to factors such as the image of computer science and computer scientists, limited knowledge about the field, and a view that computing is 'boring' [32].

Previous works have also observed significant differences in self-efficacy (one's belief in their own skills and knowledge) on all levels of schooling [12, 16, 22, 25, 35]. The gendered difference in self-efficacy is sometimes referred to as *confidence gap*, a phenomenon where women tend to rate themselves significantly lower in confidence, self-efficacy, and ability to learn and succeed when compared to men. Studies have concluded with conflicting results on whether the gap narrows with accomplishments [7] or whether the phenomenon persists, even when building experience [10].

Investigations of these phenomena are complicated by the fact that they are not constant globally. For example, investigations of STEM fields more broadly have pointed at a curious 'paradox' in that the general STEM gender gap seems to be wider in countries that are 'associated with liberal modernity' [18]. While the methodologies of some of the earlier works on this 'gender-equality paradox' have been questioned [26], further works since then have both corroborated the general finding [18] and even proposed underlying mechanisms [3].

Despite the plethora of attention, we are not aware of previous works investigating how student perceptions of what constitutes a 'good' or a 'bad' grade relates to the confidence gap, at least in the context of a Northern European university-level introductory programming course. This left us unclear to whether a part of the gap could be explained (in our context) by a situation where women – especially those who might have outperformed their peers in subjects such as mathematics on lower levels of schooling – had higher standards for what they would consider 'doing well'. Furthermore, as previous works have identified differences in self-efficacy in even nearby countries [30], and the issue is much discussed, we want to investigate whether there have been changes in the gap in the context of Finland. Motivated by these observations, this manuscript seeks to answer the following research questions:

RQ 1: How prominent is the gender confidence gap in a Finnish introductory programming MOOC in 2024?
RQ 2: How does the confidence gap evolve as students continue from an introductory programming MOOC into a subsequent advanced programming MOOC?
RQ 3: Can the gender confidence gap be explained by differences in grade perceptions?

Through these questions, we both deepen the understanding of the gender confidence gap in the context of Northern European/Nordic countries and provide an updated snapshot of how

ITiCSE 2025, Nijmegen, Netherlands

ACM ISBN 979-8-4007-1567-9/2025/06
https://doi.org/10.1145/3724363.3729103

various efforts to address the problem have (not) improved the situation.

The rest of this article is structured as follows: First, in Section 2, we provide a brief overview of some related works on gender, computer science, and the confidence gap. We then describe in Section 3 the educational context of this study, and our data collection and analysis method. The results of this analysis are then presented in Section 4. Finally, Sections 5 and 6 provide our interpretation of the results, limitations of this study, and some potential avenues of future work as well as our overall conclusions.

2 Related Work

Despite long-standing efforts, computer science remains a heavily male-dominated field in most universities. A fairly recent longitudinal study from a large computer science program in the United States reports that while the number of women majoring in computer science has increased, they still represent only some 28 % of recent majors [1]. Similarly, at the University of Helsinki, located in Finland, around 25 to 28 % of the accepted applicants to the computer science programme are women – while an improvement from the 20 % observed a decade ago, the gender balance remains very skewed.[1]

This gender imbalance in computer science studies has negative effects on women's sense of belonging, which, in turn, can affect academic performance and persistence, as well as career motivation [14]. While both male and female students in computer science experience uncertainty about belonging [14], the effects are more prominently reported by women due to, for example, negative stereotyping related to competence and lack of similar peers and role models [14, 21, 36]. In work life, gender imbalance can further exacerbate the lack of belonging, hindering, for example, women's willingness to negotiate their salaries [28]. Gender imbalances have been noted, for a very long time [13], to introduce a gendered bias to the developed tools and technologies [20]. In academic settings, though Cohoon et al. [8] report that women's authorship in computer science articles rose from 7 % in 1967 to 27 % in 2009, a more recent article by Wang et al. [33] states that the growth seems to be stagnating or even decreasing. Lack of belonging remains significant even for women who persist in academia [11].

Attempts to alleviate the issues of lack of sense of belonging for women in computer science have been made, with efforts such as increasing the mentions and visibility of female professionals in materials [21], providing role models [34], applying more gender-neutral interests to examples and exercises [5, 34], and allowing for anonymity while commenting and answering on online course platforms [15, 27]. While anonymity typically increases female students' interactions, it also prevents visible representation of gender, which, in turn, does not help to alleviate negative stereotyping or bring visible representation that could help with the sense of belonging [15, 29].

One of the suggested reasons for the continuous underrepresentation of women in computer science is the *confidence gap*, that is, women reporting lower self-confidence and self-efficacy than men. The phenomenon has been reported in, for example, computing aptitude [2] and belief in performance in programming courses [9, 24]. Just like the gender imbalance in computer science [18], the confidence gap in computing abilities is not a global phenomenon [7]. Even within the Nordic countries, Tømte and Hatlevik [30] found that from survey results collected alongside PISA[2] 2006 assessment, "Norwegian women reported higher levels of technological self-efficacy than Norwegian men, while Finnish men reported higher levels of self-efficacy than Finnish women." Additionally, the confidence gap has been shown to narrow over time [7].

The confidence gap is evident in most of the STEM fields, but is especially prominent in computer science [6]. Some studies indicate that despite the differences in previous computing experience, women exhibit more uncertainty and lack self-efficacy at the very beginning of their computer science studies. Murphy et al. [23] report in a multi-institutional study conducted in the US that although female students start their studies with less programming experience on average, they start to catch up in skill measured by grade already during the introductory courses, and are fully comparable to their male student peers by the end of Bachelor's degree. However, this study did not investigate student retention, so with the somewhat alarming dropout rates of early computer science courses [17], it is possible that the cohort at the end of the course or Bacherlor's programme is vastly different from the one in the beginning.

Similar results about prior experience in computer usage and students' comfort in using technology have been found in secondary education [25]: while female students report less experience and self-confidence in computer science topics than male students, by the end of the course, this gap narrows considerably. Indeed, the confidence gap is often visible at very early stages – Kallia et al. [16] report that even in K-12, on average, boys are more confident in their ability to succeed in computing tasks, while girls tend to underestimate their skills and ability to learn computing in the first place.

Regardless, several studies have indicated that there are no statistically significant differences between genders in programming course *performance* [4, 12, 19, 35], even when a confidence gap can be measured [12, 35]. In cases where a performance gap exists, it is usually explained by a lack of previous programming or computing experience, at least partially [4, 9]. Moreover, while feelings of success may alleviate some of the effects of the confidence gap, lower sense of belonging, gendered stereotyping, and the resulting problems with self-efficacy are still prevalent with women in computer science [6].

3 Context and method

The present study was conducted in the context of two university-level programming courses provided by University of Helsinki, a research-oriented university in Finland. The courses are attended by a variety of students ranging from computer science majors to other university students to complete outsiders taking the courses as free MOOCs.

[1]Relevant statistics only available in Finnish: https://www.helsinki.fi/fi/hakeminen-ja-opetus/hae-kandi-ja-maisteriohjelmiin/tilastoja-opiskelijavalinnoista

[2]https://www.oecd.org/en/about/programmes/pisa.html

The two courses, Introduction to Programming and Advanced Course in Programming – both worth of 5 ECTS credits[3] and in the Finnish language – are presented to the students as a larger unit, with a shared online learning material consisting of 7+7 sections, each section roughly corresponding to one week's work as expected from a CS major. The course material consists of an online textbook with several hundred embedded programming assignments, each graded automatically. The first three sections have the students work in-browser, while the rest of the assignments are completed in the VSCode editor via a plugin[4] [31].

Both courses were running continuously during this study with no deadlines and with all assignments available from the start of the study. Students could continue from the first into the second at their own pace, and even skip the introductory part (first seven sections) if they chose to and had the skills required for the advanced part. Those taking the courses as MOOCs do so without first enrolling with the university. Instead, after completing 25 % of assignments associated with each section of a course, they can take an online exam. If they pass the exam, they can formally enroll in the University of Helsinki Open University, thus allowing their studies to be registered.

As is the standard in Finland, the coursework of the participants enrolled in either via the university or the open university is graded using a six-point scale: 0 (Fail), 1 ('Passable'), 2 ('Satisfactory'), 3 ('Good'), 4 ('Very good') and 5 ('Excellent'). The grading is *not* done on a curve.

An invitation to take part in the study by filling out a questionnaire form was embedded in the material corresponding to the first section of both the Introductory and the Advanced Course. The questionnaire first provided the participants with an explanation of the study and how their data would be processed if they took part. It then asked for consent to use their responses together with data from the course for this study. We then asked the students for a user ID associated with the online course, as well as some background information, namely age (in ranges), gender, student status and whether they were CS majors, and their perception of their previous success in mathematics and technical (for example physics, chemistry, and computer science) subjects. We then asked for information about their previous programming experience (whether they had studied programming before, whether they had programmed as a hobby, and whether they had programmed in the context of their work) as well as their agreement (on a 7-step Likert scale) with a statement that they would do well on this course. Finally, we explained the university grading scale and asked for their view on what would constitute a 'good' grade on their own part, what grade they would be content with, what grade they would consider 'bad', and what grade they would expect to receive from the instant course. All questions, other than the consent, were optional.

The students were provided with one exercise point for submitting the questionnaire form. Participation in the study was fully voluntary, as the students were explicitly told that they could receive the point by simply denying the consent and leaving the rest of the questions unanswered, as long as they provided the user ID. They were also free to not interact with the questionnaire in any way.

The study did not require ethical approval per local ethical guidelines, as the participants were 15 years of age or older, the study did not deviate from informed consent, nor did it involve intervening in the physical integrity of the research participants, expose them to exceptionally strong stimuli, involve risk of causing mental harm that exceeds the limits of normal daily life, or involve a threat to the safety of participants or researchers or their family members or others close to them.

As preprocessing, we normalized the responses to the questions about grade perceptions and expectations, which were provided by the participants as free text. For both the question about what constituted a 'good' grade and what grade they would be content with, if the participant provided a range of grades, the responses were normalized to the lower provided grade (e.g. responding to '*What grade would you be content with?*' with '4-5' would be coded as 4), while the question on a 'bad' grade was coded to indicate the highest grade they would consider 'bad'.

4 Results

4.1 Participants

Prior to data analysis, we removed all responses that left some of the grade-related fields empty, participants who either declined to provide a gender[5] or self-identified as something else than 'man' or 'woman' (due to the small number of such responses, statistical analysis was not feasible for these participants), as well as participants who responded using an obviously different grading scale. This left us with N_{intro} = 391 participants from the Introductory Course. The same process applied to the data from the Advanced Course provided us N_{adv} = 127 participants. The age distributions of both populations are shown in shown in Table 1.

In the context of both courses, most respondents were not majoring in computer science. Of the responses received from Introduction to Programming participants, 28 % (57 out of 204) of the men and 14 % (26 out of 186) of the women majored in CS either at the University of Helsinki or some other institute. The corresponding numbers for the Advanced Course in Programming are 39 % (27 out of 69) of the men and 26 % (15 out of 58) of the women.

On the Introductory Course, 47 % (96 out of 205) of the men and 21 % (39 out of 186) of the women indicated having previously programmed as a hobby, with 10 % (21 out of 203) of the men and 4 % (8 out of 184) of the women indicated having programmed for work. For the Advanced Course, 51 % (35 out of 69) of the men and 50 % (29 out of 58) of the women reported having programming as a hobby, with 12% (8 out of 69) of the men and 5 % (3 out of 58) of the women having programmed for work.

The genders report broadly equivalent prior exposure to programming education. In the case of the Introductory Course, 37 % (76 out of 204) of the men and 30 % (56 out of 186) of the women had studied programming on a dedicated course, irrespective of the education level; 17 % (34 out of 204) of the men and 16 % (29 out of

[3]One ECTS credit corresponds to approximately 27 hours of work, meaning one course accounts to about 135 hours of work.

[4]https://github.com/rage/tmc-vscode

[5]In the Finnish language, the same term translates to both the English 'gender' and 'sex'. Outside of a strictly biological (and mostly non-human) context, both 'male' and 'man' would also translate to the same word in Finnish. The same applies to the Finnish equivalent of 'female' and 'woman'. The rest of this manuscript uses 'gender', 'man', and 'woman' as translations of the survey instrument's terms.

Course	Gender	15-17	18-20	21-25	26-30	31-35	36-40	41-50	51-60	61 or more
Introduction	Men	5	31	46	29	34	20	31	5	4
	Women	1	25	53	28	23	19	20	15	2
Advanced	Men	2	15	15	8	6	6	10	4	3
	Women	1	4	21	10	6	5	8	3	0

Table 1: Distribution of participant ages

186) of the women had studied programming as part of some other course (e.g. high school mathematics); and 27 % (96 out of 204) of the men and 54 % (101 out of 186) of the women had not studied programming previously. On the Advanced Course, 49 % (34 out of 69) of the men and 41 % (24 out of 58) of the women had studied on a programming course (not including the introductory course analyzed in this paper); 17 % (12 out of 69) of the men and 10 % (7 out of 58) of the women had studied programming as part of some other course; and 33 % (23 out of 69) of the men and 47 % (27 out of 58) of the women had not studied programming before.

Finally, we asked the participants about their (self-perceived) previous performance in "mathematical and technical subjects (physics, chemistry, computer science etc.)". Neither course showed statistically significant differences between the gender groups on this question, with both the Introductory Course (MWU U1=19566.0, p=0.50) and the Advanced Course (MWU U1=2312.0, p=0.10) having both groups responded with a median of 'Good'.

We return to the similarities and differences between the groups and the courses in the Discussion.

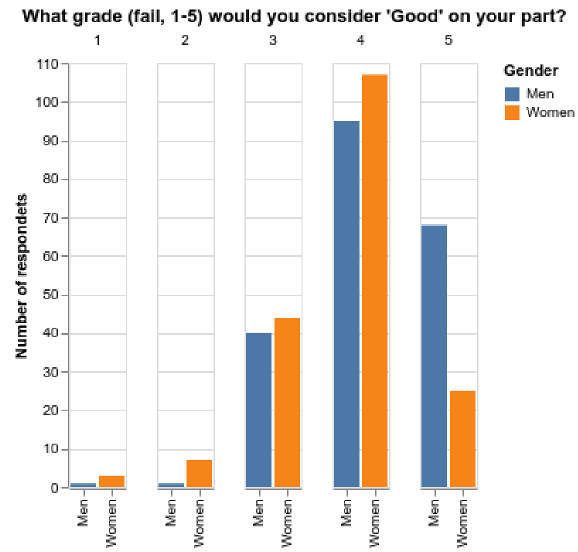

Figure 1: Responses to the question "What grade (fail, 1-5) would you consider 'good' on your part?" from the Introductory Course survey. No responses were recorded for 'fail'. The difference between the two gender groups is statistically significant (Mann-Whitney U test U1=23298.5, p=0.00003).

4.2 Views on 'good' and 'bad' grades

The responses from the Introduction to Programming participants show a small but statistically significant difference (Mann-Whitney U test U1=23298.5, p=0.00003) in terms of what grades they would view as 'good' between the men (median 4, mean 4.11, std. 0.76) and the women (median 4, mean 3.77, std. 0.79). This distribution is shown in more detail in Figure 1.

Similarly, we observe a minor but statistically significant difference (MWU U1=22340.0, p=0.0019) in terms of what grade the respondents viewed as 'bad' between the men (median 2, mean 2.08, std. 1.02) and the women (median 2, mean 1.76, std. 0.87). Finally, we observe a small difference (MWU U1=21638.0, p=0.013) between the men (median 4, mean 3.57, std. 0.88) and women (median 3, mean 3.34, std. 0.83) in terms of what grade they would be content with. While statistically significant as-is, this last p-value is no longer statistically significant if a Holm-Bonferroni correction for multiple comparisons is applied.

For the Advanced Course in Programming, we observe no gap in terms of views on what constitutes a 'good' grade (MWU U1=2018.5, p = 0.929), with both gender groups reporting a median grade of 4 (men mean 4.188, std. 0.73; women mean 4.172, std. 0.75). Similarly, we observe no gap in terms of what grade the participants report being content with (MWU U1=1931.5, p=0.724), with both gender groups reporting a median grade 4 (men mean 3.48, std. 1.02; women mean 3.57, std. 0.90). This pattern continues to the perceptions of 'bad' grades, with again no statistical difference (MWU U1=1931.0, p=0.721) as both groups report a median grade of 2 (men mean 1.97, std. 0.95; women mean 2.05, std. 0.87).

4.3 Confidence and predicted grade

When asked about their agreement with the statement that they'd perform well on the course on a 7-step Likert scale (options 'Strongly disagree', 'Disagree', 'Somewhat disagree', 'Neither disagree or agree', 'Somewhat agree', 'Agree' and 'Strongly agree', coded as a range from -3 to +3), we observe a statistically significant difference (MWU U1=23365.0, p=0.00006) between the gender groups on the Introductory Course. Men report a median value of 'Agree', while women report a median value of 'Somewhat agree'. For completeness, we also report that the groups had mean values of 1.76 and 1.35, and standard deviations of 1.10 and 1.03, respectively, but note that these values are not very meaningful for an individual Likert item. This difference disappears in the case of the Advanced Course (MWU U1=2259.5, p=0.191): both genders report a median value of 'Agree'. The equivalent means are 1.91 (std. 0.95) for men and 1.53 (std. 1.31) for women.

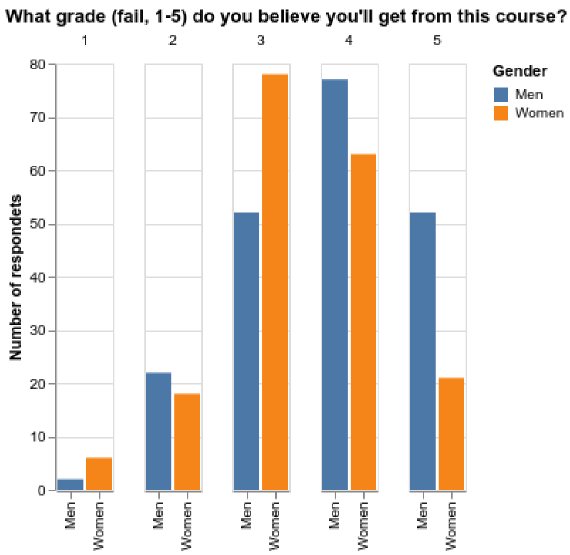

Figure 2: Responses to the question "What grade (fail, 1-5) do you believe you'll get from this course?" from the Introductory Course survey. No responses were recorded for 'fail'. The difference between the two gender groups is statistically significant (Mann-Whitney U test U1=23017.5, p=0.0002).

When asked to predict their grade as a number, for the Introductory Course we observed a statistically significant difference (MWU U1=23017.5, p=0.0002) between the gender groups. Whereas men predict a median grade of 4 (mean 3.76, std. 0.98), women predict a median grade of 3 (mean 3.40, std. 0.93). This data is shown in more detail in Figure 2. This difference, however, disappears in the case of the Advanced course in Programming (MWU U1=2092.0, p=0.642), with both groups reporting a median grade of 4 (men mean 3.88, std. 0.88; women mean 3.81, std. 0.85).

4.4 Descriptions of predicted grades

Next, we computed the difference between the self-reported 'good' grade and the grade the respondents anticipate receiving, e.g. denoting with '+1' a case where a respondent indicates they believe they will receive the grade 5, and would consider the grade 4 'good'. For the Introductory Course in Programming, we observe no statistically significant (MWU U1=18512.5, p=0.58) difference between the men (mean 0.36, std. 0.69) and the women (mean 0.37, std. 0.67). Along the same lines, there was no statistically significant difference in the delta between the 'bad' grade and the predicted grade (MWU p=0.20; men mean -1.67, women mean -1.63) or the grade one would be content with and the predicted grade (MWU p=0.48; men mean -0.19, women mean -0.06). The same pattern is present in the data for the Advanced Course, with no statistically significant difference between genders in terms of the delta between the expected and the 'good' grade (MWU U1=1985.5, p=0.935; men mean 0.30, std. 0.67; women mean 0.36, std. 0.77), the delta between the expected and the grade the respondents would be content with (MWU U1=17786.0, p=0.20; men mean -0.185, std. 0.70; women mean -0.06, std. 0.80), or the delta between the expected and the 'bad' grade (MWU U1=18350.0, p=0.482; men mean -1.67, std. 0.81; women mean -1.64, std. 0.75).

4.5 Evolving confidences

As the two courses are both constantly running, data was collected from both concurrently and reporting of identifiers that allowed merging the datasets was optional, we only have data from 28 individuals where we can conclusively identify their responses to both questionnaires. Of these 28, 19 are men and 9 are women. The limited data makes statistical analysis difficult. We calculated the between-courses deltas of what the participants viewed as 'good' (mean 0.07, std. 0.54) and 'bad' (mean -0.14, std. 0.70) grades, what grades they would be content with (mean -0.14, std. 0.59), what grade they predicted for themselves (mean -0.07, std. 0.72), and their agreement with the 'will perform well' question (mean -0.11, std. 0.83). None of these showed a statistically significant difference between the two gender groups, with the smallest observed p-value for the Mann-Whitney U test being p=0.411 for change in what was considered the 'bad' grade.

4.6 Course progress

As noted previously, the main matter of the courses consisted of more than a hundred individual exercises, with a fully optional final exam. As only a fraction of the participants would take the exam, we instead investigate how many points (i.e. completed exercises) they received by the data collection cut-off date. Notably for this analysis, the data collection date was not the end date of the course, and thus the points analyzed here do not necessarily mean the final point totals the participants would go to achieve, but rather only represent a lower bound.

In the case of the Introductory Course, we observe that men received (from parts associated with that course only) a median point total of 88.0 (mean 94.3, std. 72.7), while the women's median was 68.5 (mean 77.3, std. 65.0), which is a statistically significant difference using the standard threshold (MWU, U1=21509.0, p=0.028), but fails to achieve statistical significance if we correct for multiple comparisons using Holm-Bonferroni. For the Advanced Course (again, looking only at points relating to that course), the differences are not statistically significantly different even prior to applying a correction for multiple comparisons (MWU U1=2081.0, p=0.670), with men attaining a median of 99.0 points (mean 90.4, std. 54.2) and women having a median of 94.5 points (mean 89.1, std. 51.7).

We also analyzed the number of material sections from which the students had attained any points. For the Introductory Course, men obtained points from a median of 7 sections (i.e. all; mean 5.09, std. 2.29), while women obtained points from a median of 5 sections (mean 4.62, std. 2.39). This difference is barely statistically significant prior to applying a correction for multiple comparisons (MWU U1=21188.5, p=0.0412), but fails to reach the Holm-Bonferroni corrected threshold. For the Advanced Course, we observe no statistically significant difference (MWU U1=2242.5, p=0.212) between the men (median 7.0, mean 5.36, std. 2.22) and the women (median 6.0, mean 5.00, std. 2.20).

Finally, we investigated whether there were differences in how the respondents to the Advanced Course questionnaire fared in the Introductory Course in terms of points obtained. We observed no statistically significant difference (MWU U1=2087.0, p=0.676) between the men (median 172, mean 133.6, std. 79.5) and the women (median 168, mean 126.5, std. 82.6). We will return to this result in particular in the Discussion.

5 Discussion

Returning to our research questions, our data leads us to the following answers:

RQ 1: How prominent is the gender confidence gap in a Finnish introductory programming MOOC in 2024? We observed a statistically significant (MWU U1=23365.0, p=0.00006, men median 'Agree', women median 'Somewhat agree') difference between the gender groups in terms of agreement with a claim that the respondent would do well on the Introductory Course. We also found a statistically significant difference in what the gender groups predicted as their concrete grade (MWU U1=23017.5, p=0.0002, men median 4, women median 3). We interpret this as a strong signal that the gender confidence gap still exists in university-level CS1 education in Finland, despite long-standing efforts to reduce or eliminate the effect. These findings replicate results from other studies in similar contexts [2, 9, 24].

RQ 2: How does the confidence gap evolve as students continue from an introductory programming MOOC into a subsequent advanced programming MOOC?. Our analysis failed to identify the same gender confidence gap in the data for the Advanced Course using both the agreement question and the predicted grade. On the surface, this could be viewed as a signal that the confidence gap disappears as students are exposed to programming. However, our data also indicates that there are material differences in the women respondents between the two courses. For one, the percentage of women who had programmed as a hobby rose from 21 % on the Introductory Course to 50 % on the Advanced Course, while the equivalent numbers for men were a more consistent 47 % and 51 %. Furthermore, we observed a nearly statistically significant difference between the men's and the women's point totals obtained from the Introductory Course participants (MWU U1=21509.0, p=0.028; median 88.0, 68.5; does not reach a Holm-Bonferroni corrected threshold), but the Introductory Course point totals for those responding to the Advanced Course survey were much closer (MWU U1=2242.5, p=0.212; median 172 vs. 168). Further data and study is needed, but we hypothesize that a major contributor in the "disappearance" of the gender confidence gap is actually the dropping out of low-confidence women. This could be at least partially supported by previous findings that indicate that women may adjust their self-efficacy beliefs earlier than men [19], meaning that early difficulties with programming may cause them to drop out more easily.

RQ 3: Can the gender confidence gap be explained by differences in grade perceptions? We found no evidence to support this hypothesis. If anything, our data points towards the opposite on the Introductory Course: there is a statistically significant albeit small difference between grade expectations in terms of what constitutes a 'good' grade (MWU U1=23298.5, p=0.00003) and what constitutes a 'bad' grade (MWU U1=22340.0, p=0.0019). In both cases, men's responses have a higher means than the women's responses at 4.11 vs. 3.77 and 2.08 vs. 1.76, respectively. This difference is not identifiable from the data for the Advanced Course, possibly relating to the above hypothesis on the women population meaningfully changing due to women dropping out more than men.

As always, the above results need to be understood in their context. Importantly, our data contains responses from a mixture of people taking the courses as part of their CS major, those taking them as a minor subject, and those taking them as a free MOOC course. As the amount of statistical analyses was already relatively large in comparison to the respondent population, we elected not to investigate differences between these groups in the present work. Second, as our data comes from a Finnish language course, and thus we do not expect a meaningful number of international participants, these results should not be necessarily interpreted as extending beyond the context of Finnish students. Indeed, previous works have shown differing results between even nearby and closely related countries [30]. Furthermore, some of our analyses produced p-values that reached the traditional threshold of statistical significance (0.05), but not a threshold that corrects for multiple comparisons using the Holm-Bonferroni method. We believe these results should be investigated further using another – potentially larger – dataset.

Our results suggest a need to look deeper into the degree to which changing student perceptions are caused by changing student beliefs (e.g. increasing self-efficacy and confidence on the part of women students), and to what degree they are caused by changing student populations (e.g. low-self-efficacy and low-confidence students dropping out). In the future, we intend to explore this further in the context of our CS programme as a whole, rather than only individual courses.

6 Conclusion

We set out to provide an updated snapshot of the status of the gender confidence gap in CS1 education in the Northern European context, as well as to test whether the gender confidence cap could be in part explained by women demanding more of themselves than men. Our results indicate that the gender confidence gap lives on in the Finnish context despite long-standing efforts to reduce or eliminate it completely. Our data also led us to a hypothesis that the 'improvement' observed in the confidence gap as studies progress could be explained (at least in part) by low-self-efficacy and low-confidence women dropping out, rather than their self-efficacy and confidence improving. Further work is needed to investigate this hypothesis. Our results also indicate that the gender confidence gap – at least in the context of Finland – is not explainable by women having higher expectations in terms of e.g. what numeric grade they consider 'good'. In fact, our data suggests the opposite, even if the effect is small in magnitude.

References

[1] Amy Baer and Andrew DeOrio. 2020. A Longitudinal View of Gender Balance in a Large Computer Science Program. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 23–29. https://doi.org/10.1145/3328778.3366806

[2] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. 2003. Gender differences in computer science students. In *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education* (Reno, Navada, USA) *(SIGCSE '03)*. Association for Computing Machinery, New York, NY, USA, 49–53. https://doi.org/10.1145/611892.611930

[3] Thomas Breda, Elyès Jouini, Clotilde Napp, and Georgia Thebault. 2020. Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences* 117, 49 (2020), 31063–31069.

[4] Amy Bruckman, Carlos Jensen, and Austina DeBonte. 2023. Gender and programming achievement in a CSCL environment. In *Computer Support for Collaborative Learning*. Routledge, 119–127.

[5] Sapna Cheryan, Victoria Plaut, Paul Davies, and Claude Steele. 2009. Ambient Belonging: How Stereotypical Cues Impact Gender Participation in Computer Science. *Journal of personality and social psychology* 97 (12 2009), 1045–60. https://doi.org/10.1037/a0016239

[6] Sapna Cheryan, Sianna Ziegler, Amanda Montoya, and Lily Jiang. 2016. Why are some STEM fields more gender balanced than others? *Psychological Bulletin* 143 (10 2016). https://doi.org/10.1037/bul0000052

[7] MacKenzie A. Christensen. 2023. Tracing the Gender Confidence Gap in Computing: A Cross-National Meta-Analysis of Gender Differences in Self-Assessed Technological Ability. *Social Science Research* 111 (2023), 102853. https://doi.org/10.1016/j.ssresearch.2023.102853

[8] J. McGrath Cohoon, Sergey Nigai, and Joseph "Jofish" Kaye. 2011. Gender and computing conference papers. *Commun. ACM* 54, 8 (Aug. 2011), 72–80. https://doi.org/10.1145/1978542.1978561

[9] Rodrigo Duran, Lassi Haaranen, and Arto Hellas. 2020. Gender Differences in Introductory Programming: Comparing MOOCs and Local Courses. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 692–698. https://doi.org/10.1145/3328778.3366852

[10] Jessica Ellis, Bailey K. Fosdick, and Chris Rasmussen. 2016. Women 1.5 Times More Likely to Leave STEM Pipeline after Calculus Compared to Men: Lack of Mathematical Confidence a Potential Culprit. *PLOS ONE* 11, 7 (07 2016), 1–14. https://doi.org/10.1371/journal.pone.0157447

[11] Katrina Falkner, Claudia Szabo, Dee Michell, Anna Szorenyi, and Shantel Thyer. 2015. Gender Gap in Academia: Perceptions of Female Computer Science Academics. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education* (Vilnius, Lithuania) *(ITiCSE '15)*. Association for Computing Machinery, New York, NY, USA, 111–116. https://doi.org/10.1145/2729094.2742595

[12] Brian Harrington, Shichong Peng, Xiaomeng Jin, and Minhaz Khan. 2018. Gender, confidence, and mark prediction in CS examinations. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (Larnaca, Cyprus) *(ITiCSE 2018)*. Association for Computing Machinery, New York, NY, USA, 230–235. https://doi.org/10.1145/3197091.3197116

[13] Charles Huff and Joel Cooper. 1987. Sex Bias in Educational Software: The Effect of Designers' Stereotypes on the Software They Design. *Journal of Applied Social Psychology* 17, 6 (1987), 519–532. https://doi.org/10.1111/j.1559-1816.1987.tb00328.x

[14] Elisabeth Höhne and Lysann Zander. 2019. Sources of Male and Female Students' Belonging Uncertainty in the Computer Sciences. *Frontiers in Psychology* 10 (2019). https://doi.org/10.3389/fpsyg.2019.01740

[15] David A. Joyner, Lily Bernstein, Ian Bolger, Maria-Isabelle Dittamo, Stephanie Gorham, and Rachel Hudson. 2022. Anonymity: A Double-Edged Sword for Gender Equity in a CS1 Forum?. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for Computing Machinery, New York, NY, USA, 766–772. https://doi.org/10.1145/3478431.3499289

[16] Maria Kallia and Sue Sentance. 2018. Are boys more confident than girls? the role of calibration and students' self-efficacy in programming tasks and computer science. In *Proceedings of the 13th Workshop in Primary and Secondary Computing Education* (Potsdam, Germany) *(WiPSCE '18)*. Association for Computing Machinery, New York, NY, USA, Article 16, 4 pages. https://doi.org/10.1145/3265757.3265773

[17] Päivi Kinnunen and Lauri Malmi. 2006. Why students drop out CS1 course?. In *Proceedings of the Second International Workshop on Computing Education Research* (Canterbury, United Kingdom) *(ICER '06)*. Association for Computing Machinery, New York, NY, USA, 97–108. https://doi.org/10.1145/1151588.1151604

[18] Seungah S Lee, Christine Min Wotipka, Francisco O Ramirez, and Jieun Song. 2024. To STEM or not to STEM: A cross-national analysis of gender and tertiary graduates in science, technology, engineering, and math, 1998–2018. *International Journal of Comparative Sociology* (2024), 00207152241243343.

[19] Alex Lishinski, Aman Yadav, Jon Good, and Richard Enbody. 2016. Learning to Program: Gender Differences and Interactive Effects of Students' Motivation, Goals, and Self-Efficacy on Performance. In *Proceedings of the 2016 ACM Conference on International Computing Education Research* (Melbourne, VIC, Australia) *(ICER '16)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/2960310.2960329

[20] J. Margolis and A. Fisher. 2002. *Unlocking the Clubhouse: Women in Computing*. MIT Press.

[21] Allison Master, Sapna Cheryan, and Andrew Meltzoff. 2015. Computing Whether She Belongs: Stereotypes Undermine Girls' Interest and Sense of Belonging in Computer Science. *Journal of Educational Psychology* 108 (08 2015). https://doi.org/10.1037/edu0000061

[22] Martina RM Meelissen and Marjolein Drent. 2008. Gender differences in computer attitudes: Does the school matter? *Computers in Human behavior* 24, 3 (2008), 969–985.

[23] Laurie Murphy, Brad Richards, Renée Mccauley, Briana Morrison, Suzanne Westbrook, and Timothy Fossum. 2006. Women catch up: Gender differences in learning programming concepts. *SIGCSE Bulletin* 38, 17–21. https://doi.org/10.1145/1121341.1121350

[24] Nea Pirttinen, Arto Hellas, Lassi Haaranen, and Rodrigo Duran. 2020. Study Major, Gender, and Confidence Gap: Effects on Experience, Performance, and Self-Efficacy in Introductory Programming. In *2020 IEEE Frontiers in Education Conference (FIE) (Frontiers in Education Conference)*. IEEE, United States. https://doi.org/10.1109/FIE44824.2020.9273884

[25] Yolanda Reimer and Michael Coe. 2022. A Narrowing of the CS Knowledge and Confidence Gap in High School Students by Gender. In *Proceedings of Society for Information Technology & Teacher Education International Conference 2022*, Elizabeth Langran (Ed.). Association for the Advancement of Computing in Education (AACE), San Diego, CA, United States, 1621–1626. https://www.learntechlib.org/p/220925

[26] Sarah S Richardson, Meredith W Reiches, Joe Bruch, Marion Boulicault, Nicole E Noll, and Heather Shattuck-Heidorn. 2020. Is there a gender-equality paradox in science, technology, engineering, and math (STEM)? Commentary on the study by Stoet and Geary (2018). *Psychological Science* 31, 3 (2020), 338–341.

[27] Naaz Sibia, Angela Zavaleta Bernuy, Joseph Jay Williams, Michael Liut, and Andrew Petersen. 2023. Student Usage of Q&A Forums: Signs of Discomfort?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 33–39. https://doi.org/10.1145/3587102.3588842

[28] Adina D. Sterling, Marissa E. Thompson, Shiya Wang, Abisola Kusimo, Shannon Gilmartin, and Sheri Sheppard. 2020. The confidence gap predicts the gender pay gap among STEM graduates. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30303–30308. https://doi.org/10.1073/pnas.2010269117

[29] Adrian Thinnyun, Ryan Lenfant, Raymond Pettit, and John R. Hott. 2021. Gender and Engagement in CS Courses on Piazza. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 438–444. https://doi.org/10.1145/3408877.3432395

[30] Cathrine Tømte and Ove E. Hatlevik. 2011. Gender-differences in Self-efficacy ICT related to various ICT-user profiles in Finland and Norway. How do self-efficacy, gender and ICT-user profiles relate to findings from PISA 2006. *Computers & Education* 57, 1 (2011), 1416–1424. https://doi.org/10.1016/j.compedu.2010.12.011

[31] Arto Vihavainen, Thomas Vikberg, Matti Luukkainen, and Martin Pärtel. 2013. Scaffolding students' learning using test my code. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education* (Canterbury, England, UK) *(ITiCSE '13)*. Association for Computing Machinery, New York, NY, USA, 117–122. https://doi.org/10.1145/2462476.2462501

[32] Anna Vitores and Adriana Gil-Juárez. 2016. The trouble with 'women in computing': a critical examination of the deployment of research on the gender gap in computer science. *Journal of Gender Studies* 25, 6 (2016), 666–680.

[33] Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, and Oren Etzioni. 2021. Gender trends in computer science authorship. *Commun. ACM* 64, 3 (2021), 78–84.

[34] Kelly Widdicks, Alice Ashcroft, Emily Winter, and Lynne Blair. 2021. Women's Sense of Belonging in Computer Science Education: The Need for a Collective Response. In *Proceedings of the 2021 Conference on United Kingdom & Ireland Computing Education Research* (Glasgow, United Kingdom) *(UKICER '21)*. Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3481282.3481288

[35] Chris Wilcox and Albert Lionelle. 2018. Quantifying the Benefits of Prior Programming Experience in an Introductory Computer Science Course. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 80–85. https://doi.org/10.1145/3159450.3159480

[36] Angela Zavaleta Bernuy, Anna Ly, Brian Harrington, Michael Liut, Sadia Sharmin, Lisa Zhang, and Andrew Petersen. 2023. "I Am Not Enough": Impostor Phenomenon Experiences of University Students. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 313–319. https://doi.org/10.1145/3587102.3588779