An Investigation of Gender Differences in Computer Science Using Physiological, Psychological and Behavioural Metrics

Keith Nolan
Maynooth University
Kildare, Ireland
keith.nolan@mu.ie

Aidan Mooney
Maynooth University
Kildare, Ireland
aidan.mooney@mu.ie

Susan Bergin
Maynooth University
Kildare, Ireland
susan.bergin@mu.ie

ABSTRACT

Gender imbalance in tertiary Computer Science (CS) and Information Technology (IT) courses is a cause for concern globally. Current estimates of this imbalance are ~70:30 male to female. Within the CS education field numerous studies have investigated the cause of this imbalance (e.g. misconceptions of CS, stereotypes, etc.) and have attempted to identify factors that influence uptake, retention, and performance. Whilst these studies have had varying degrees of success, none appear to have investigated in-the-moment physiological differences between genders during module examinations. Such research could provide new insight on how male and female students process and respond in such a setting and provide new opportunities to better tailor module delivery and assessment.

This paper describes a novel study that investigates gender differences in skin conductance and heart rate variability during a controlled exam-like setting. Participant background information such as gender, age, previous experience, etc. was collected at the outset. The examination was designed and validated in-house using a peer-review process and carefully constructed to ensure that only one new concept was introduced per question. General behavioural metrics, such as doodling, response time, and researcher observations were gathered. An out-of-the-box psychological test was used to measure self-reported anxiety and physiological arousal was measured using wearable sensors before and during the experiment. Study design, methodology, and analysis are described in detail. Findings suggest differences exists between genders in physiological and behavioural responses when completing programming comprehension questions. The findings provide valuable evidence to justify future research in this area.

CCS CONCEPTS

• **Social and professional topics → CS1**;

KEYWORDS

Gender, physiological signals, CS1, programming

ACM Reference Format:
Keith Nolan, Aidan Mooney, and Susan Bergin. 2019. An Investigation of Gender Differences in Computer Science using Physiological, Psychological and Behavioural Metrics. In *Twenty-First Australasian Computing Education Conference (ACE'19), January 29–31, 2019, Sydney, NSW, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3286960.3286966

ACE'19, January 29–31, 2019, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6622-9/19/01…$15.00
https://doi.org/10.1145/3286960.3286966

1 MOTIVATION

Over the years many initiatives have taken place to attempt to bridge the gender gap in the uptake of CS with varying degrees of success. It appears that little research has been conducted on the physiological differences between genders while completing programming tasks. This paper aims to highlight these differences and inform the community on the possible impacts that different questions (concepts) could have on the different genders.

There is a considerable disparity in the number of female students choosing to study Computer Science compared to male students at tertiary education level [3, 4, 14, 31]. In the 1980's, Computer Science had one of the highest rates of gender balance in graduate programmes, but this has changed considerably in recent times [37]. At our university, a gender imbalance is evident from first to final year. First year Computer Science modules and in particular CS1, tend to be modules that students on many different degree streams take to make up course credits. The current gender split is ~80:20 male to female. By contrast, final year Computer Science is taken only by students who wish to finish with a Computer Science qualification. Final year Computer Science tends to be highly male-dominated. In 2008 there was a gender split of 78% male and 22% female in final year Computer Science at our university. However, by 2018 this split had increased to 89% male and 11% female. This is not unusual. The Higher Education Authority in Ireland released a report in 2016 documenting a country-wide gender difference in the field of Computer Science and Maths, with a split of 81% male and 19% female [14].

Further evidence from around the world shows similar trends. In the USA, the percentage of females pursuing a degree in Computer Science has gone from 40% in 2000–2001 to 26% in 2008–2009. The percentage of women receiving degrees in Computer Science is even lower, with a 28% completion rate in 2000 compared to 17.7% in 2008 [30]. This number further decreased in 2011 with the release of the Computer Research Association report stating that less than 12% of Computer Science degrees were awarded to women [27]. The U.S. Department of Commerce Economics and Statistics Administration released a report outlining that only 27% of the workforce in Computer Science and math were women [2]. More recently, in 2017, the U.S. Department of Labour reported that only 25.5% of people working in the Computer Science and IT field were women [23]. Similar trends were noted in the UK with the WISE Campaign citing only 14% of the ICT workforce are women [38].

Over the years many initiatives have tried to address the downward trend in female participation. As such, this paper is presented as follows. Related literature is described in Section 2 followed by the research questions addressed in this study in Section 3. The setup and methodology are detailed in Sections 4 and 5. Detailed findings are outlined in Section 6 followed by threats to validity, conclusions and future work.

2 RELATED LITERATURE

Different initiatives have been trialled over the years to combat the downward trend in female participation in CS education. Of note, the European Union (EU) funded one of the largest recent initiatives aimed to encourage females to participate in STEM subjects across many EU countries. A website entitled: "Science: It's a girls thing!" aimed at teenage girls was developed and included information on careers within the STEM fields and a quiz to discover their "inner researcher". Accompanying the website was a video that depicted female "scientists" conducting work in stiletto heels. This video was referred to as offensive and after criticism was removed [31]. Other initiatives such as *Girls Who Code*, *SciGirls* and *GirlsInc* have been created. Indicative reports suggest that they are improving the uptake of females in CS, however, no formal studies have been conducted.

Female perception of the STEM field can be negative with views that CS is a "nerdy" and male-dominated subject. Several reasons have been cited for this including CS majors being deficient in interpersonal skills and male tutors/educators displaying a superiority complex [4]. Several studies have interviewed women to better understand what it is like to be a woman in Computer Science. Of note was that respondents indicated that they could not see the point of coding and that they preferred to code alone at home and not in a lab as they did not feel like they belonged there [24]. In another study, female respondent's indicated that they felt uncomfortable with assistance given to them by male tutors [30].

Considerable research has been carried out on confidence and self-efficacy in Computer Science. Females display significantly lower confidence and programming self-efficacy in Computer Science when compared to males [1, 3, 18, 27, 31]. This is concerning as programming self-efficacy is significantly correlated to success in CS1 [26]. A recent large-scale study, involving 690 students across 11 different institutions, examined perceived self-efficacy and test anxiety during a programming exam [26]. Findings from this study indicated significant differences in the self-efficacy and test anxiety of genders within CS1. An interesting finding from the study was that males tended to outperform females at the early stage of CS1. However, in the later stages of CS1 females tended to outperform males. Quille et al. suggested that this difference may be caused by females having lower programming self-efficacy than males [26]. The study also found that females have greater test anxiety and that this may affect performance. While the direction of performance is not noted in the study, Nunez-Pena et al. conducted a study to investigate if self-reported test anxiety differences between genders has any correlation with academic performance [22]. Findings suggested that there were no correlations between genders in self-reported test anxiety and academic performance. However, findings by Nunez-Pena et al. suggest that females do, in general, have greater levels of test anxiety, which is in line with Quille et al. These findings are further supported by Cassady et al. who suggest that the gender difference is due to females appraising the test situation as more threatening than males [8].

Nolan et al. recently carried out an extensive systematic literature review of the role of anxiety in CS with an emphasis on learning to program [21]. Findings from this review suggest that students are anxious when learning to program resulting from a multitude of factors, including, task complexity, the modality of programming assessments and general anxiety of using a computer. While studies have investigated if anxiety played a role in computer programming, very few have examined any inherent gender differences of that anxiety. New studies that attempt to understand and address the gender imbalance in CS are crucial and related literature indicates that studies that examine gender differences with respect to exam anxiety are worthy of pursuit.

With physiological sensors becoming more common in everyday life, for example, heart rate sensors in smartwatches, activity trackers, and mobile devices, the opportunity to monitor in-the-moment physiological changes in a learning setting is considerably more accessible. Recent related research has shown that wearable technologies can be used in the classroom to aid and possibly improve teaching by using the technology to monitor engagement to the course material [25, 29]. Similarly, McNeal et al. [20] investigated the use of Electrodermal Activity (EDA) over the course of a lecture to measure the engagement of a class during three key parts: the lecture, the movie and the dialogue. Seventeen students participated in the study. These students wore an EDA sensor for the duration of the class and filled out pre-class and post-class questionnaires on the teaching concept. Findings from the study showed that; a) students were more engaged during the movie than any other part of the class, and, b) the higher the emotional arousal of the participants, the more engaged they were with the material.

More recently Photoplethysmogram (PPG) has been used in the detection of emotional arousal in learning. A recent study [12] examined students heart rate during a 50-minute lecture. As the lecture progressed, the heart rate of students decreased until the point where one student asked a question. It was then observed that the heart rates of all the students increased.

Building on this recent work, the aim of this study is to investigate gender differences during a controlled lab exam using physiological metrics, as well as standard psychological and behavioural metrics.

3 RESEARCH QUESTIONS

This paper investigates the following research questions:

(1) Are there differences in the physiological signals (Electrodermal Activity, Heart Rate Variability) between male and female students during an MCQ in a controlled lab setting?
(2) Are there differences in the behavioural activity (doodling, researcher observations and response time) between male and female students during an MCQ in a controlled lab setting?
(3) Are there differences in State and Trait anxiety between male and female students in CS1?

4 INSTRUMENTS

4.1 Background Survey

As part of this experiment, a short background questionnaire for participants was developed that contained items relating to the participant's age, gender and other attributes pertaining to the setup of the experiment, such as their dominant hand.

The State-Trait Anxiety Inventory (STAI) was used to gather self-assessed anxiety level prior to the experiment [32]. The STAI is routinely used as a clinical survey in diagnosing anxiety and is arguably the most commonly used tool in the evaluation of anxiety with 12 language versions available. The survey contains 40 questions, 20 relating to State anxiety and 20 relating to Trait anxiety. The survey is graded on a 4-point likert scale and has been extensively used to assess levels of state anxiety which have been induced by a stressful experiment. The higher the overall score the greater the level of anxiety.

State anxiety is defined as an unpleasant emotional arousal in the face of threatening demands or dangers. A cognitive appraisal of threat is a prerequisite for the experience of this emotion [16]. Trait anxiety refers to the tendency to attend to, experience, and report negative emotions such as fears, worries, and anxiety across many situations [13]. The higher the Trait anxiety measure, the more susceptible you are to experience anxiety, i.e. someone with high trait anxiety might respond negatively to a stimulus whereas someone with low trait anxiety possibly would not respond at all.

As the experiment took place towards the end of CS1, we had to retrospectively collect the final module result for the participants after they had completed the final module exam. This allowed us to profile the students based on performance.

4.2 Physiological Sensors

Participants wore two physiological sensors for the duration of the experiment, an Electrodermal Activity (EDA) sensor and a Photoplethysmography (PPG). Both sensors were part of the Shimmer 3 GSR+, which is a wireless device used to record EDA and PPG signals. Both the EDA and PPG were sampled at 51.2Hz to ensure we captured subtle changes in the signals. The Shimmer was placed on the wrist of the non-dominant hand to ensure the participants could use the mouse and doodle. The PPG sensor was placed on the tip of the middle finger with the EDA sensors placed on the tips of the index and ring fingers of the non-dominant hand.

Electrodermal Activity is one of the most commonly used measures for physiological response, with studies focusing on a variety of tasks from measuring attention to predicting abnormal behaviours [6]. EDA sensors work by measuring the resistance between two points on the skin. EDA is used in this study to determine when a student becomes aroused, that is when a student begins to react to a situation and a physiological response (heart or sweat rate increases) occurs if that arousal is constructive or destructive.

A photoplethysmography is a sensor that detects changes in the volume of blood flow by measuring the difference in the light reflected into the sensor. Using various algorithms, the heart rate can be estimated along with other measures such as Heart Rate Variability from the PPG values. The PPG was used to capture the heart beat-to-beat data and from this, the Heart Rate Variability (HRV) could be calculated. HRV, the beat to beat variations of the heart rate, is a known indicator of the interplay between the sympathetic and parasympathetic nervous system. This interplay can provide an insight into the fight-flight response in the body. If there is more variability in the beat to beat data, the person could have perceived a threat, and so the systems that controls the beating of the heart begin to fight for control. Using this knowledge, HRV measures can be used as an indicator of emotional arousal [17].

4.3 Programming Comprehension Questions

4.3.1 Java Course. The students who took part in this study were taking the CS1 module in our university. CS1 is our introduction to computer programming module. Students in this module have no previous formal study of programming. The module runs over twelve weeks and consists of three hours of lectures and three hours of labs per week. The module covers programming fundamentals, typically delivered in the following order: variables, types, expressions and assignment; simple I/O, Conditional and iterative control structures (if statements and while loops), Strings and string processing, arrays and other fundamentals such as problem-solving and computer architecture. This module structure is similar to the proposed Java Programming 1 course outlined in the ACM Curriculum Guidelines for Undergraduate Programs in Computer Science [11].

4.3.2 Development of questions. The programming comprehension exam was designed in-house and based on the Java programming language. Each question was subject to the following constraints:

- Multiple choice in nature
  - there were four possible answers,
  - there was only one correct answer, and,
  - one "None of these" answers.
- Always had a clear output i.e. there was no hidden challenges or tricks in the question.

To allow for fine-grained analysis each question contained only one new concept, for example: a loop, a conditional statement or a string. This allowed the responses to be analysed both individually and collectively so that the most likely cause(s) of difficulty could be identified. In addition to this, as physiological data was being collected throughout the experiment, if any changes in the physiological signals were detected we could correlate them to a specific concept.

4.3.3 Validation. Thirteen questions were developed in total. Initially, the questions were reviewed by the authors to categorise their difficulty level as either easy, medium, or hard. To further ensure that the questions developed were of good and sound quality seven postgraduate research students (Research Masters/PhD candidates) in the Computer Science department in our university were recruited to:

(1) Review all questions to get a sense of the range of concepts asked,
(2) Answer each question to ensure that the correct answer was identifiable, and finally,
(3) Rate each question on a scale of 1 (easy) to 9 (hard) in terms of difficulty.

Table 1: Core concepts shown in the experiment

Concepts	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
System output	New	x	x	x	x	x	x	x	x
String variables		New	x						x
String Concatenation			New						
If—else if—else statements				New					
Nested if-else statements					New				
While Loop						New	x		
Substring							New	x	
If—else with Substring								New	
Nested While Loop									New

Results from the surveys showed that of the initial 13 questions, all the postgraduate students' ratings were similar across all questions. Average rating scores were taken, and the questions were ranked in order of difficulty based on said scores. One question was removed as one postgraduate student got the question wrong and so it may have been too hard for novice learners. Two questions were removed as they were "too long" compared to the other questions and so would not fit on the presentation screen. One question that had been labelled by the authors as "easy" was labelled as "medium" by several of the postgraduate reviewers. This question was removed. This resulted in 9 peer-validated questions (3 at each difficulty level) for the experiment.

Each question builds on the previous question with the first being the easiest and the ninth being the hardest, thus, the first three questions (Q1, Q2, Q3) were considered easy, the middle three questions (Q4, Q5, Q6) were considered medium, and the last three questions (Q7, Q8, Q9) were considered hard. Table 1 shows the nine core concepts examined in the experiment. The table also outlines the concept that was new, depicted by the word "New" for that particular question and what other concepts were contained in each question (by use of x). These core concepts were chosen as they were the concepts that the participants would have been exposed to in the CS1 course.

As this experiment took place late in the CS1 course, participants had more exposure to certain concepts that were covered earlier in the module. For example, the participants would be very familiar with system output as they would use this in (virtually) every program that they write. Comparing this to concepts such as substring or nested loops, participants would have only been introduced to them towards the end of CS1. Following the validation of the questions, the expectation was that all participants should get the easy questions correct, most would get the medium questions correct and only some would get the hard questions correct.

5 DESIGN

5.1 Participants

Participants were studying CS1 in our university and were gathered on a voluntary basis. Ethical approval was sought and granted to carry out this research.

5.2 Experimental Protocol

One researcher and participant were alone in the room. The researcher was out of view from the participant throughout the experiment and stayed in the room to ensure the experiment ran smoothly. Participants were instructed to read an information sheet describing the experiment prior to commencement. If they had any issues or questions they were encouraged to ask for clarification. Once completed, they were asked to sign a consent form. The background survey and STAI were then given to the participant for completion.

The physiological sensors were then placed on the non-dominant hand of the participant. A short 30-second baseline measurement was taken at the beginning of the experiment to ensure the sensors were functioning and the participant was comfortable with the sensors. During this baseline measurement, the Shimmer 3 GSR+ was calibrated using it's on board software. The participant was seated and encouraged to stay as relaxed as was possible.

After obtaining the baseline measurement, the participants started the MCQ test. All questions were presented evenly counterbalanced in groups of Easy, Medium or Hard. Within each difficulty band, questions were always shown in the same order as outlined in Section 4.3.3. This was done to ensure that there was no confounding effect.

The participant was instructed to answer each question by using the mouse to click on the answer they thought was correct. Each participant was provided with a pen and paper and told they were allowed to doodle. The participants were asked by the conducting researcher to think and consider each question carefully. Observations were noted by the researcher throughout the experiment.

6 ANALYSIS

6.1 Participant Profile

Forty-two participants (30 male, 12 female) participated in this study. Table 2 presents the age and gender profiles of the participants.

Table 2: Age and gender profile of participants

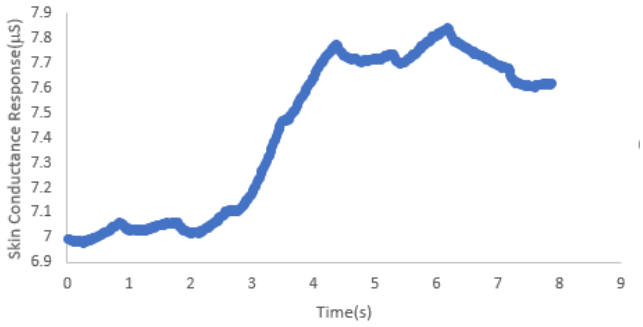
Age	Male (N=30)	Female (N=12)
17–19	22 (74%)	10 (84%)
20–22	4 (13%)	1 (8%)
23+	4 (13%)	1 (8%)

6.2 Research Question 1: Are there differences in the physiological signals (Electrodermal Activity, Heart Rate Variability) between male and female students during an MCQ in a controlled lab setting?

Electrodermal Activity

To investigate emotional arousal an algorithm created by Taylor et al. in MIT Media Lab was used that automatically identifies Skin Conductance Response (SCR) [34]. An SCR is the change in the Skin Conductance Level that may be associated with a stressor and an increase in sweat production. Figure 1 shows a snippet of a participants SCR. After identifying all SCRs, any participant that had a number of SCRs outside two standard deviations of the mean was removed as they were considered outliers within the data. This resulted in the removal of two males and one female.

Figure 1: Skin Conductance Response



We compared the number of SCRs over the entire experiment and across the question difficulty band. A high number of SCRs indicated a high level of emotional arousal. Table 3 describes the gender breakdown of SCRs.

Table 3: Average number of SCRs across the experiment and across each question difficulty band categorised by gender

	Male (N=28)	Female (N=11)	p-value
Total Average	16.64	8.72	0.056
Easy	2.6	1.16	**0.044**
Medium	6.03	2.25	**0.018**
Hard	9.96	5.33	0.11

As can be seen in Table 3 there is a gender difference in the number of SCRs over the course of the programming exam. Male participants demonstrated almost double the number of SCRs throughout the experiment compared to female participants, however as will be discussed in Section 6.3, males did not outperform female students. A Welch's t-test was used to compare the data to a 95% significance level. This returned a p-value of 0.056. This value is tending towards significance and the result is still valuable especially given the small size of the participant pool.

To further examine the gender difference in EDA, we reviewed the number of SCRs over the difficulty bands. Table 3 provides the number of SCRs across question difficulty band. A Welch's t-test was used to compare the data to a 95% significance level.

Similar to SCRs across the entire experiment, males tended to be significantly more aroused throughout all difficulty bands of the programming comprehension questions. Further to this, if we observe the increase in SCRs across the difficulty bands, the number of female SCRs increased the most between Medium and Hard whereas male SCRs increased the most between Easy and Medium.

While no studies have investigated the difference in physiological signals between genders in an exam-like scenario, other studies in different disciplines (public speaking, psychological tests and responses to music) have found that females have higher EDA readings [7, 28] or have found no discernible difference [15]. Although the settings/focus are very different this is interesting to note, and further study is justified.

Photoplethysmography

A PPG was used during the experiment to capture the heart beat-to-beat data. From this data, factors such as heart rate and the Root Mean Square of the Successive Differences (RMSSD) could be calculated. The RMSSD is a measure of HRV and has been found to correlate to emotional arousal [9].

The beat-to-beat data was cleaned using a low pass filter to reduce the noise and sharpen the beat-to-beat peaks. Following this, a peak detection algorithm was used to create interbeat intervals (IBI) which is the time between successive beats of the heart. Finally, customised software was developed in-house to analyse the IBI file and determine the RMSSD. The PPG signals were analysed over the first and second halves of the experiment. We chose to analyse over halves rather than individual questions or question bands as the time frame for reliable HRV measures may be too short. Halves were chosen by taking the total run time of the experiment and dividing by 2. One set of participant data had to be removed from the PPG data as sections of the data was lost during recording. Table 4 presents the gender difference in PPG data on average over the course of the experiment.

Table 4: Average ln(RMSSD) values for each experiment half categorised by gender

	Male	Female	p-value
Half 1 ln (RMSSD)	4.892	5.252	0.089
Half 2 ln (RMSSD)	4.913	5.221	0.118

From Table 4 we can see that there is a difference in the PPG data across the experiment. The natural log (ln (RMSSD)) of the RMSSD scores was used to compare across participants. In both halves females have a higher ln (RMSSD) score. This is an indication of a higher heart rate throughout the experiment. Table 4 also shows the significance values when compared across the genders over the experiment.

While there are no statistical differences between the genders there is a numerical difference with male participants having lower heart rate variability scores. Comparing normal figures found in

HRVCourse.com, HRV levels for this age group are typically significantly lower, approximately 24% lower than what was observed in the study. Although males typically have lower HRV scores than females [35], the observed increase in HRV scores could be attributed to test-anxiety given the exam-like situation that the experiment was carried out in. Further research is required on a larger population with an even gender balance to investigate why the HRV scores are higher than normal.

It is important to note that gender differences do exist in both the EDA and HRV. Typically, females have higher EDA and HRV scores than males. This study found that males have higher EDA scores than female and this requires further investigation. A large-scale recent study found that female students outperformed male students at the late stages of CS1 [26]. At the time this experiment was conducted, students had completed eight of twelve weeks of the CS1 module. Given that female students perform better in the latter stages of CS1, they could ultimately be more confident in their ability and may not be getting stressed or aroused when presented with the programming questions and thus produce fewer SCRs than male students. A follow-up study focused on this hypothesis is needed.

6.3 Research Question 2: Are there differences in the behavioural activity (doodling, researcher observations and response time) between male and female students during an MCQ in a controlled lab setting?

During the experiment we collected information on the questions participants doodled on, the researcher's observations and the response time for each question. We found that in general female students were significantly faster and more accurate when completing these programming MCQ questions.

Researchers Observations

During the experiment, the conducting researcher noted any behavioural issues that might impact the validity of the results. Eight observations were noted in total. These were:

- Three participants laughed or made a comment during the experiment,
- One participants' phone rang,
- One participant broke their glasses before the experiment (they reported that they still were able to see the screen),
- One participant had a shaking hand throughout the experiment,
- One participant spent particularly long on one question, and,
- One participant was anxious before starting the experiment.

No significant differences in observations were noted between male and female participants.

Response Time

When investigating the differences in response time, we examined only the response times of the correct responses. Table 5 outlines these response times and significant values per question.

A Welch's t-test was used to compare the data to a 95% significance level. Breaking the response times down on a per question level, females, on average, responded faster than males and in some cases were significantly faster. Every participant had to respond to each of the questions. There was no time limit enforced on answering the questions as we wanted to ensure that the participant answered the questions as accurately as possible. Although a significant difference was found on five out of the nine questions, it is important to draw attention to question 9 (focused on nested loops) — only a small number of female participants correctly answered this question. This low number may call the validity of the value of that finding into question.

Table 5: Percentage of correct answers, average time taken in seconds (s) to respond correctly to each question, grouped by gender

	% Correct		Response Time (s)		
	Male	Female	Male	Female	p-value
Q1	100%	100%	14.98	13.79	0.307114
Q2	96.67%	91.67%	14.78	11.63	**0.0265**
Q3	70%	66.67%	23.39	15.49	**0.0398**
Q4	100%	100%	22.05	16.15	**0.01735**
Q5	83.33%	100%	26.65	18.68	**0.002928**
Q6	56.67%	50%	50.85	59.49	0.273882
Q7	23.33%	33.33%	110.81	69.97	0.1584
Q8	56.67%	58.33%	41.03	36.05	0.28744
Q9	46.67%	33.33%	79.1	45.01	0.0549

Table 6: Participant breakdown in exam performance bands

	Male (N=30)	Female (N=12)
Top Performers (70%+)	53%	50%
Middle Performers (40%-69%)	37%	42%
Low Performers (<40%)	10%	8%

Exam bands

As the results for the end of module exam were collected, we were able to band participants into three main profiles: top performers—participants who achieved 70%+ in their final CS1 exam, middle performers—participants who achieved 40%–69% in their final CS1 exam and low performers—participants who achieved <40% in their final CS1 exam. Table 6 shows the breakdown of participants and gender in each of these bands.

It is important to note that top performers account for 52% of our participants. This shows that the sample population is not representative of the class and results may be biased by this.

Top Performers

Top performers were those students who achieved 70%+ in the final module exam. The response time for correct answers was reviewed. Table 7 presents the differences that exist in top performance band along with the significance values.

Of the nine questions in the experiment, two significant differences exists within the top performers when response time was examined. Both Question 4 and 5 examined versions of *If statements* and on both questions, females responded faster. This suggests that females at a top performer level may have a better understanding of the concept and so can respond faster when compared to males. In

addition to this, no females got Question 9 correct and therefore no significance tests could be run and so Question 9 was excluded from Table 7. Question 6 examined a *While Loop* and Question 9 examined *Nested loops*. While there is no significant difference, females performed worse than males on Question 6 and 9. Loops often cause an issue for CS1 students, however this notable gender-related difference is very interesting and a future study, with a larger number of female participants, should be carried out to re-examine this finding.

Middle Performers

The middle performers were also examined to see if any gender difference existed. Table 8 shows the differences in the middle performance band. The response time from only the correct responses was analysed.

As can be seen from Table 8, several questions had significant differences and on almost all questions, females were faster and more accurate when responding to the questions. As no participant in this performance band got Question 7 correct, it was excluded from the table.

Low Performers

In the low performers band there were three males and one female. No significance tests could be run given the low number of participants in the band.

Doodling

Doodling/code tracing when completing programming questions is often encouraged in lab situations with educators urging students to write down on paper what they want to do before they code it. A study conducted by Lister et al. [19] showed that those who doodle/code trace tend to perform better. As part of this experiment we encouraged participants to doodle if they needed to.

Participants were provided with a pen and blank paper. The sheets were collected after the experiment and doodles categorised as to what question they related to. Table 9 outlines the gender differences in doodles on a per question basis, with the significance levels included. This analysis was done on a binary level, either the participant doodled on a question or they did not. In order to determine significance levels, a Welch's t-test was conducted. It is important to note that no one doodled on Questions 1, 2 or 8. There were significant differences between male and female participants doodling on Questions 3 and 5. However, this should be interpreted with caution as no female participant doodled on Questions 3, 4 and 5. Excluding these questions, no significant differences were observed and this conforms to the literature that no gender difference exists [5].

Table 7: Average time taken in seconds to respond to each question categorised by gender for the top performers

	Male Average	Female Average	p-value
Question 1	14.096	17.102	0.2335
Question 2	15.585	13.382	0.1789
Question 3	23.452	15.652	0.1439
Question 4	19.941	13.243	**0.0254**
Question 5	24.703	14.72	**0.0004**
Question 6	48.851	60.912	0.2516
Question 7	110.82	69.977	0.1585
Question 8	34.668	32.136	0.3652

Table 8: Average time taken in seconds to respond to each question categorised by gender for the middle performers

	Male Average	Female Average	p-value
Question 1	15.885	11.061	**0.019**
Question 2	14.167	8.391	**0.005**
Question 3	24.139	16.831	0.123
Question 4	25.324	18.906	0.116
Question 5	29.386	23.068	0.128
Question 6	60.565	56.668	0.463
Question 8	50.425	25.578	**0.044**
Question 9	100.794	37.843	0.082

Table 9: Percentage breakdown of doodles per question with significance values

	# of Doodles	Male (N=30)	Female (N=12)	p-value
Q1	0	0 (0%)	0 (0%)	-
Q2	0	0 (0%)	0 (0%)	-
Q3	2	2 (100%)	0 (0%)	0.0804
Q4	1	1 (100%)	0 (0%)	0.1628
Q5	3	3 (100%)	0%	**0.0415**
Q6	17	12 (70%)	5 (30%)	0.4624
Q7	15	11 (73%)	4 (27%)	0.4223
Q8	0	0 (0%)	0 (0%)	-
Q9	11	7 (63%)	4 (37%)	0.2728

Discussion

In general females were significantly faster and more accurate than males when completing the programming comprehension questions at this stage of the module; this is in line with previous research [26]. Interestingly, with respect to accuracy and response time within the exam performance bands (Low, Middle and Top performers), there are significant differences in both Top and Middle performance bands, with females in both bands tending to respond faster than males.

In addition, female participants doodled less than their male counterparts. Females outperformed males on Question 5 (nested if statements) as outlined in Table 5, however, no female doodled on this question. Perhaps an interesting insight is the fact that the top performing female students under performed on the loop questions and there is an increase in the number of female doodles on these questions. This might suggest that loops are a stumbling block for female students and warrants further investigation. While it is important to point out that this study examined just the frequency of doodles on a per question level, the quality of the doodle is just as important. Research conducted by Cunningham et al. and Whalley et al. examined the types and quality of doodles made by CS1 programmers [10, 36]. Findings presented by Cunningham et al. and Whalley et al. suggest that those that doodle tend to be more

successful when completing a programming question. While the finding presented above is contradictory to the literature, a future aspect of this study could be to investigate the types and quality of the doodles made by CS1 students with a larger population.

6.4 Research Question 3: Are there differences in State and Trait anxiety between male and female students in CS1?

All participants completed the STAI at the start of the experiment and results were calculated using the scoring key provided with the STAI. Normalised results were obtained from the STAI manual. Table 10 outlines the normalised average values of both self-reported State and Trait anxiety.

As can be seen in Table 10, there is no significant difference between male and female participants. Both male and female State and Trait averages appear similar. This finding is consistent with the normative averages in the STAI averages for College Students [33]. The State averages for both male and female participants are within normal limits. Interestingly, the Trait levels are considerably higher than the normal values as outlined in the manual.

Table 10: Average State and Trait values for male and female participants and associated p-values

Age Profile	Male	Female	p-value
State	38.96	40.25	0.44
Trait	52	56.83	0.32

Previously, data captured from a large multi-institutional international study found that females have greater test anxiety than males [26]. In this study, while there was no significant difference between the genders in both State and Trait anxiety, females exhibit marginally higher anxiety levels. This, however, is normal as per the STAI manual.

Of interest is the considerably higher than normal levels of Trait anxiety in both male and female participants in this study. The Trait averages for male and female students in a normal college population are 38.30 and 40.40 respectively. There is, on average a 15.065 point increase in self-reported Trait anxiety than what is reported in the STAI manual. Given that we know Computer Science students are anxious [21], this higher than normal Trait anxiety suggests that the students who participated in this study are slightly more prone to perceive situations as more stressful than what they are. As was discussed in Section 6.3, our sample population is biased towards top performers with 52% of participants achieving 70%+ in the final module exam. Given this bias and higher than normal Trait anxiety, further research is required to investigate if there is any correlation between Trait anxiety and success in CS1.

7 THREATS TO VALIDITY

While every precaution was taken when designing and setting up this study, the following outlines some potential limitations:

(1) There appears to be a slight disconnect between the researcher's evaluation of difficulty (easy, medium and hard) in the programming questions when compared to the success rate on the questions.
(2) As noted in the study, the participants who took part in the study were biased towards the top of the class. The sample of females was representative of the female cohort in general, however, the male sample was slightly biased towards the top of the class.
(3) This experiment was carried out with just the participant and the researcher in the room, a further study would be valuable in a more authentic exam-like setting.
(4) The baseline was taken at rest for a 30-second period at the beginning of the experiment. To ensure a more accurate baseline measurement, ideally, the participants would wear a sensor in the days leading to the experiment. This would provide a more stable baseline.

8 CONCLUSIONS AND FUTURE WORK

This study set out to examine if gender differences existed within the CS1 community from three different perspectives: psychological, physiological and behavioural. In the psychological measures, self-reported State and Trait anxiety, there is very little difference in the self-reported data between the genders. We observed normal State anxiety scores and slightly increased Trait anxiety scores. These close to normal scores may be related to the voluntary nature of participants, with perhaps some students choosing not to participate because of anxiety. The considerably higher self-reported State and Trait anxiety scores for both males and females in CS1 is significant and warrants further investigation. Developing ways to mitigate a student's anxious predispositions while teaching would be a valuable addition.

The most novel part of this study was the use of physiological metrics. Examining the physiological measures during a programming exam-like situation allows us a unique insight into how the students are interacting with the questions. The study found that there are differences in both the EDA and HRV readings. While known gender differences already exist in the readings, both EDA and HRV scores were higher than normal. With wearable technology becoming more accessible, educators could potentially use physiological signals in their class to track arousal of their students while teaching a particular concept.

Given the observed differences between genders across the psychological, physiological and behavioural metrics further research is justified. Exploring the difference metrics for female students we observed an increase in SCRs in the hard questions and in the top performing female students we observed a decrease in the number of correct responses to loop related questions. We also observed an increase in doodling from female students in the loop questions. The observed differences for female students with loop-related questions warrant follow up study a to better interpret the findings and establish their importance. A repeat study, ideally, with a higher number of female participants should be carried out with an emphasis on physiological metrics.

9 ACKNOWLEDGEMENTS

The authors would like to thank the participants of this study. This work was funded by the Irish Research Council Postgraduate Scholarship 2015 GOIPG/2015/1671.

REFERENCES

[1] Mustafa Baser. 2013. Attitude, Gender and Achievement in Computer Programming. *Middle-East Journal of Scientific Research* 14, 2 (2013), 248–255.

[2] David N Beede, Tiffany A Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E Doms. 2011. Women in STEM: A gender gap to innovation. https://files.eric.ed.gov/fulltext/ED523766.pdf

[3] Sylvia Beyer, Michelle DeKeuster, Kathleen Walter, Michelle Colar, and Christina Holcomb. 2005. Changes in CS Students' Attitudes Towards CS over Time: An Examination of Gender Differences. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '05)*. ACM, New York, NY, USA, 392–396. https://doi.org/10.1145/1047344.1047475

[4] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. 2003. Gender Differences in Computer Science Students. *SIGCSE Bull.* 35, 1 (Jan. 2003), 49–53. https://doi.org/10.1145/792548.611930

[5] Jason Bruce Boggs, Jillian Lane Cohen, and Gwen C. Marchand. 2017. The Effects of Doodling on Recall Ability. *Psychological Thought* 10, 1 (2017), 206–216. https://doi.org/10.5964/psyct.v10i1.217

[6] John T Cacioppo, Louis G Tassinary, and Gary Berntson. 2007. *Handbook of psychophysiology*. Cambridge University Press.

[7] Eduvigis Carrillo, Luis Moya-Albiol, Esperanza Gonzalez-Bono, Alicia Salvador, Jorge Ricarte, and Jesus Gomez-Amor. 2001. Gender Differences in Cardiovascular and Electrodermal Responses to Public Speaking Task: The Role of Anxiety and Mood States. *International Journal of Psychophysiology* 42, 3 (2001), 253 – 264. https://doi.org/10.1016/S0167-8760(01)00147-7

[8] Jerrell C. Cassady and Ronald E. Johnson. 2002. Cognitive Test Anxiety and Academic Performance. *Contemporary Educational Psychology* 27, 2 (2002), 270 – 295. https://doi.org/10.1006/ceps.2001.1094

[9] Kwang-Ho Choi, Junbeom Kim, O. Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji-Eun Park. 2017. Is Heart Rate Variability (HRV) an Adequate Tool for Evaluating Human Emotions? - a Focus on the Use of the International Affective Picture System (IAPS). *Psychiatry Research* 251 (2017), 192 – 196. https://doi.org/10.1016/j.psychres.2017.02.025

[10] Kathryn Cunningham, Sarah Blanchard, Barbara Ericson, and Mark Guzdial. 2017. Using Tracing and Sketching to Solve Programming Problems: Replicating and Extending an Analysis of What Students Draw. In *Proceedings of the 2017 ACM Conference on International Computing Education Research (ICER '17)*. ACM, New York, NY, USA, 164–172. https://doi.org/10.1145/3105726.3106190

[11] The Joint Task Force on Computing Curricula, Association for Computing Machinery, and IEEE-Computer Society. 2013. *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. ACM, Inc. https://doi.org/10.1145/2534860

[12] Diana Darnell and Paul Krieg. 2014. Use of Heart Rate Monitors to Assess Student Engagement in Lectures. *The FASEB Journal* 28, 1 (2014), 721–25.

[13] Yori Gidron. 2013. *Trait Anxiety*. Springer New York, New York, NY, 1989–1989. https://doi.org/10.1007/978-1-4419-1005-9_1539

[14] David Harmon and Stephen Erskine. 2017. Eurostudent Survey VI. http://hea.ie/assets/uploads/2018/01/HEA-Eurostudent-Survey.pdf

[15] W. Huang and R. Benjamin Knapp. 2017. An Exploratory Study of Population Differences Based on Massive Database of Physiological Responses to Music. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 524–530. https://doi.org/10.1109/ACII.2017.8273649

[16] Theodore D. Kemper and Richard S. Lazarus. 1992. Emotion and Adaptation. *Contemporary Sociology* 21, 4 (1992), 522. https://doi.org/10.2307/2075902

[17] Richard D. Lane, Kateri McRae, Eric M. Reiman, Kewei Chen, Geoffrey L. Ahern, and Julian F. Thayer. 2009. Neural Correlates of Heart Rate Variability During Emotion. *NeuroImage* 44, 1 (2009), 213 – 222. https://doi.org/10.1016/j.neuroimage.2008.07.056

[18] 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 (ICER '16)*. ACM, New York, NY, USA, 211–220. https://doi.org/10.1145/2960310.2960329

[19] Raymond Lister, Elizabeth S. Adams, Sue Fitzgerald, William Fone, John Hamer, Morten Lindholm, Robert McCartney, Jan Erik Moström, Kate Sanders, Otto Seppälä, Beth Simon, and Lynda Thomas. 2004. A Multi-national Study of Reading and Tracing Skills in Novice Programmers. *SIGCSE Bull.* 36, 4 (June 2004), 119–150. https://doi.org/10.1145/1041624.1041673

[20] Karen S. McNeal, Jacob M. Spry, Ritayan Mitra, and Jamie L. Tipton. 2014. Measuring Student Engagement, Knowledge, and Perceptions of Climate Change in an Introductory Environmental Geology Course. *Journal of Geoscience Education* 62, 4 (Nov 2014), 655–667. https://doi.org/10.5408/13-111.1

[21] Keith Nolan and Susan Bergin. 2016. The Role of Anxiety when Learning to Program: A Systematic Review of the Literature. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research (Koli Calling '16)*. ACM, New York, NY, USA, 61–70. https://doi.org/10.1145/2999541.2999557

[22] Maria Isabel Nunez-Pena, Macarena Suarez-Pellicioni, and Roser Bono. 2016. Gender Differences in Test Anxiety and Their Impact on Higher Education Students' Academic Achievement. *Procedia - Social and Behavioral Sciences* 228 (2016), 154 – 160. https://doi.org/10.1016/j.sbspro.2016.07.023 2nd International Conference on Higher Education Advances,HEAd'16, 21-23 June 2016, Valencia, Spain.

[23] Bureau of Labor Statistics. 2017. Women in the Labor Force: A Databook. https://www.bls.gov/opub/reports/womens-databook/2017/home.htm

[24] Reena Pau, Wendy Hall, Marcus Grace, and John Woollard. 2011. Female Students' Experiences of Programming: It's Not All Bad!. In *Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education (ITiCSE '11)*. ACM, New York, NY, USA, 323–327. https://doi.org/10.1145/1999747.1999837

[25] M. Poh, N. C. Swenson, and R. W. Picard. 2010. A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity. *IEEE Transactions on Biomedical Engineering* 57, 5 (May 2010), 1243–1252. https://doi.org/10.1109/TBME.2009.2038487

[26] Keith Quille, Natalie Culligan, and Susan Bergin. 2017. Insights on Gender Differences in CS1: A Multi-institutional, Multi-variate Study. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '17)*. ACM, New York, NY, USA, 263–268. https://doi.org/10.1145/3059009.3059048

[27] Katie Redmond, Sarah Evans, and Mehran Sahami. 2013. A Large-scale Quantitative Study of Women in Computer Science at Stanford University. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education (SIGCSE '13)*. ACM, New York, NY, USA, 439–444. https://doi.org/10.1145/2445196.2445326

[28] Sonja Rohrmann, Henrik Hopp, and Markus Quirin. 2008. Gender Differences in Psychophysiological Responses to Disgust. *Journal of Psychophysiology* 22, 2 (2008), 65–75. https://doi.org/10.1027/0269-8803.22.2.65 arXiv:https://doi.org/10.1027/0269-8803.22.2.65

[29] Brian K Sandall. 2016. Wearable Technology and Schools: Where are We and Where Do We Go From Here? *Journal of Curriculum, Teaching, Learning and Leadership in Education* 1, 1 (2016), 9.

[30] Donna Saulsberry. 2012. Dwindling Number of Female Students: What Are We Missing?. In *Proceedings of the 13th Annual Conference on Information Technology Education (SIGITE '12)*. ACM, New York, NY, USA, 221–226. https://doi.org/10.1145/2380552.2380616

[31] Jane Sinclair and Sara Kalvala. 2015. Exploring Societal Factors Affecting the Experience and Engagement of First Year Female Computer Science Undergraduates. In *Proceedings of the 15th Koli Calling Conference on Computing Education Research (Koli Calling '15)*. ACM, New York, NY, USA, 107–116. https://doi.org/10.1145/2828959.2828979

[32] Charles D Spielberger. 1989. *State-Trait Anxiety Inventory: Bibliography (2nd ed.)*. CA: Consulting Psychologists Press.

[33] Charles D Spielberger, R. Gorsuch, R. L.and Lushene, P. R. Vagg, and G. A Jacobs. 1983. *Manual for the State-Trait Anxiety Inventory*. CA: Consulting Psychologists Press.

[34] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard. 2015. Automatic Identification of Artifacts in Electrodermal Activity Data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 1934–1937. https://doi.org/10.1109/EMBC.2015.7318762

[35] Andreas Voss, Rico Schroeder, Andreas Heitmann, Annette Peters, and Siegfried Perz. 2015. Short-Term Heart Rate Variability-Influence of Gender and Age in Healthy Subjects. In *PloS one*.

[36] Jacqueline Whalley, Christine Prasad, and P. K. Ajith Kumar. 2007. Decoding Doodles: Novice Programmers and Their Annotations. In *Proceedings of the Ninth Australasian Conference on Computing Education - Volume 66 (ACE '07)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 171–178. http://dl.acm.org/citation.cfm?id=1273672.1273693

[37] Brenda Wilson. 2008. Improving Comfort Level of Females in the First Computer Programming Course: Suggestions for CS Faculty. *J. Comput. Sci. Coll.* 23, 4 (April 2008), 28–34. http://dl.acm.org/citation.cfm?id=1352079.1352086

[38] WISE. 2017. Women in STEM workforce 2017. https://www.wisecampaign.org.uk/statistics/women-in-stem-workforce-2017/