Female Students Assess Software Tutors More Positively Than Male Students

Amruth N. Kumar
Ramapo College of New Jersey, amruth@ramapo.edu

Abstract - In an earlier study reported at FIE 2006, we found that there was no statistically significant difference between female and male students on their assessment of the usability, ability to learn from and the usefulness of our software tutors, except on the only two negatively worded Likert-scale statements with which female students disagreed significantly more than male students. In order to find out whether this difference was simply an artifact of the negative wording of the statements or whether female students assess software tutors more positively than male students, we conducted a follow-up study using a feedback form that contained an even number of positively- and negatively-worded Likert-scale statements. We found that female students agreed with positively-worded statements and disagreed with negatively-worded statements significantly more than male students. We conclude that female students assess our software tutors more positively than male students and this is not an artifact of the (positive versus negative) wording of the feedback statements.

Index Terms – Assessment of software tutors, Feedback forms, Gender differences.

INTRODUCTION

Over the last two decades, many researchers have studied gender differences in introductory Computer Science in order to explain the phenomenon of the shrinking pipeline [1]. In the process, they have found many differences between men and women, some significant and others insignificant, with some differences still persisting while others have gradually diminished.

For instance, female students typically come to introductory Computer Science course with less prior programming experience than do men [2]. Fortunately, there is no correlation between prior experience and success in the major as measured by the grade point average at graduation [3].

In the nineties, researchers found a gender divide: women wanted to do useful things with computing while men liked to focus on programming and the machine itself. Researchers summarized this as “computing with a purpose” versus “dreaming in code” [2]. After increasing enrollment of women and providing support services for women such as professional, networking, and mentoring opportunities [4], researchers at the same institution found that men and women liked programming equally, and they did not relate to the discipline differently, i.e., in terms of programming versus application as had been observed before, even though the curriculum had not been modified in any way to become “female-friendly” [5].

Female students have lower self-confidence than male students in computer-related abilities [2,6,7]. Women have a tendency to enter computing classes with considerably less confidence than men [8]. Women are less confident than men in their ability to achieve their educational goals in computing at the undergraduate [9] and the graduate level [10]. This is true even when female students have the same level of skills as male students [11]. Fortunately or unfortunately, both male and female students have greater confidence than actual skill [11].

We have ourselves been interested in examining differences between male and female students in the context of the software tutors that we have been developing for introductory Computer Science. What we have found to date is that there is no statistically significant difference between the learning of female and male students using our software tutors [12]. Using our tutors helps improve the self-confidence of female students to be on par with that of male students [13]. Unfortunately, there is no correlation between the change in self-confidence and the change in learning of students using our tutors [14].

We wanted to find out whether women felt differently than men about using software tutors, since this would have implications for the appeal of software tutors. One early study suggests that most women who have taken online courses are generally enthusiastic about computer-based learning [15]. But, our interest was in using online tutors in traditional courses. In order to find out whether women felt differently than men, we conducted two studies in fall 2004 and spring 2005 [16] using software tutors that we have developed for computer programming. We analyzed data from the use of four tutors each semester in the introductory Computer Science course. We found no statistically significant difference between male and female students on their assessment of the tutors except on two statements that happened to be the only two negatively-worded statements on the feedback form. So, it was unclear whether female students were more positive in their assessment of our tutors, or they were less likely to agree with negative statements in general.

We conducted a follow-up study in spring 2007 using two software tutors on while and for loops. For the new study, we re-worded the statements on the feedback form so that it contained an even number of positively- and...
negatively-worded statements. We analyzed the collected data to find out whether there was any difference between male and female students and whether there was any difference between positively- and negatively-worded feedback statements. In this paper, we will describe the study — the feedback form, the data we collected and the results of analysis. In the next section, we will briefly describe our software tutors. We will summarize the results of our earlier study, followed by a description of the new study. We will end with a discussion of the results of the new study, which support the assertion that female students assess software tutors more positively than male students. Participants of our study were asked to identify their sex (biological notion of male/female) rather than their gender (social/cultural notion of man/woman) [17]. Therefore, our analysis will be in terms of sex rather than gender.

SOFTWARE TUTORS

The software tutors used for this study are web-based tutors that we have developed on programming language concepts such as loops, selection statements and expression evaluation, called probllets (www.probllets.org). The tutors are designed to help students learn programming concepts by solving problems. The tutors adapt the problems to the learning needs of the students [18]. They provide step-by-step explanation of the problem solution, which has been shown to help students learn [19]. The tutors are available for C++, Java and C# programming languages. Each tutor presents problems to the student, grades the student’s answer and provides feedback, including step-by-step explanation of the program underlying the problem. Typically, a student goes through the following stages when using a tutor:

- Registration, when the student enters demographic information;
- Pretest, during which the student’s prior level of knowledge on the topic is assessed by the tutor;
- Practice, during which the student solves problems on the concepts on which the student made mistakes during the pretest;
- Post-test, during which the student’s improvement in learning is assessed on the concepts on which the student solved practice problems;
- Feedback, during which the student enters feedback about the tutor, including the learnability of the tutor, user experience, usability of the tutor and its usefulness. Students use the tutors usually after class, on their own time. All the evaluation data is collected on-line, in-vivo and the data from the students in all the sessions that use the tutor in a semester is combined for analysis purposes.

EARLIER STUDY

At the end of each tutoring session, students are asked to fill out a feedback form. The feedback form contains 12–13 statements on a Likert scale of 1 (Strongly Agree) to 5 (Strongly Disagree). The statements were:

1. It was easy to learn how to use this tutor.
2. It was easy to learn to use this tutor.
3. The generated problems were instructive.
4. The feedback provided to my answers was clear.
5. The feedback provided to my answers was useful.
6. The feedback provided to my answers was sufficient.
7. The tutor helped clarify what I already knew.
8. The tutor helped me learn new material.
9. Using this tutor to learn was time-consuming.
10. The generated problems were repetitive and boring.
11. This tutor should be made available to all students.
12. If this tutor is made available, I will use it.
13. I would like to see such tutors on other topics.

Note that statements 1 and 2 relate to the usability, statements 3–10 relate to the ability to learn from (i.e., learnability) and statements 11–13 relate to the usefulness of the tutor. Only statements 9 and 10 were negative statements about the tutor. On these two statements, higher ratings (e.g., 5 for Strongly disagree) corresponded to more positive evaluation of our tutors, whereas on all other statements, lower ratings (e.g., 1 for Strongly agree) corresponded to more positive evaluation of our tutors.

For the previous study [16], we analyzed the combined data collected through four tutors in fall 2004 and spring 2005. In both fall 2004 and spring 2005, the difference between female and male students was greater than 20% on only two statements — 9 and 10, and all four differences were statistically significant as shown in Table 1.

<table>
<thead>
<tr>
<th>Question</th>
<th>Male Ave No. Resp.</th>
<th>Female Ave No. Resp.</th>
<th>t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>2.33 95</td>
<td>2.85 60</td>
<td>0.004</td>
</tr>
<tr>
<td>10</td>
<td>2.19 96</td>
<td>2.92 60</td>
<td>0.00007</td>
</tr>
</tbody>
</table>

These differences could mean that female students assess our software tutors more positively than male students. But, since these were the only two statements on the feedback form that were negatively worded, we could not rule out that the differences were simply an artifact of the negative wording of the two statements. Hence our new study.

NEW STUDY

For the new study conducted in spring 2007, we re-worded and re-ordered the statements on the feedback form so that it contained an even number of positively- and negatively-worded statements. We once again used the same Likert-
scale of 1 (Strongly Agree) to 5 (Strongly Disagree). The statements were:
1. The generated problems were instructive.
2. The feedback provided to my answers was NOT clear.
3. The feedback provided to my answers was useful.
4. The feedback provided to my answers was NOT sufficient.
5. The tutor helped me learn new material.
6. Using this tutor to learn was time-consuming.
7. The generated problems were repetitive and tiring.
8. The progress of my learning was NOT presented clearly.
9. It was easy to use this tutor.
10. It was NOT easy to learn how to use this tutor.
11. It was clear to me after each problem, how much I knew and how much I had yet to learn.
12. This tutor should be made available to all the students.
13. If this tutor is made available, I would NOT use it.
14. I would like to see such tutors on other topics.

Note that statements 1, 3, 5, 9, 11, 12 and 14 are positively-worded and the rest are negatively-worded. Statements 1–5 relate to the ability to learn from the tutor, statements 6, 7, 8 and 11 relate to user experience, statements 9–11 relate to usability and statements 12–14 relate to the usefulness of the tutor.

We collected data from the two loop tutors (for and while) in spring 2007. We combined the data from the students of 10 sections that used the while loop tutor and 9 sections that used the for loop tutor.

I. while Loop Tutor Results

First, we analyzed data to find out whether there was any difference between positively- and negatively-worded statements. We did a univariate ANOVA with feedback response as the dependent variable and sex as the fixed factor. We found a significant interaction between sex and the type of feedback statement (positive versus negative) as fixed factors. Once again, we found a significant interaction between sex and the type of feedback statement (positive versus negative) as fixed factors. We did a univariate ANOVA with feedback response as the dependent variable and sex and the type of feedback statement (positive versus negative) as fixed factors.

We “normalized”, i.e., reversed the Likert scale on negative feedback statements so that a score of 1 meant positive assessment on all the 14 statements and re-analyzed the feedback data. We found a significant main effect for sex [F(1,1994) = 7.311, p = 0.007] – female students agreed significantly more with the feedback statements than male students (average of 2.64 versus 2.77). In other words, they evaluated the tutors significantly more positively than male students.

Next, we analyzed the feedback by groups: the ability to learn from the tutor (statements 1–5), user experience (6–8, 11) usefulness of the tutor (statements 9–10) and usability of the tutor (statements 12–14). We used normalized feedback for these analyses since the number of positively and negatively worded statements was not even under the four categories. We did a univariate ANOVA with normalized feedback response as the dependent variable and sex as the fixed factor.

Finally, we analyzed the responses on each statement one by one. In Table II, we have listed the normalized averages for male and female students on the 14 statements – note that female responses are almost always more positive than male responses. However, we did not find a significant main effect for sex on any of the statements.

II. for Loop Tutor Results

Once again, we started by analyzing the difference between positively- and negatively-worded statements. We did a univariate ANOVA with feedback response as the dependent variable and sex and the type of feedback statement (positive versus negative) as fixed factors. Once again, we found a significant interaction between sex and the type of statement [F(1,1359) = 12.895, p = 0.000]:

- On learnability, we did not find a significant main effect for sex [F(1,707) = 1.124, p = 0.289].
- On user experience, we found a significant main effect for sex [F(1,568) = 4.946, p = 0.027]: female students rated their experience significantly more positively than male students (average of 3.053 versus 3.267).
- On usefulness, we did not find a significant main effect for sex [F(1,1286) = 0.142, p = 0.707].
- On usability, we did not find a significant main effect for sex [F(1,430) = 2.498, p = 0.115].

We collected data from the two loop tutors (for and while) in spring 2007. We combined the data from the students of 10 sections that used the while loop tutor and 9 sections that used the for loop tutor.

**I. while Loop Tutor Results**

First, we analyzed data to find out whether there was any difference between positively- and negatively-worded statements. We did a univariate ANOVA with feedback response as the dependent variable and sex and the type of feedback statement (positive versus negative) as fixed factors. We found a significant interaction between sex and the type of statement [F(1,1994) = 7.311, p = 0.007]:

- Female students agreed with positive statements more than male students (average of 2.37 versus 2.47, F(1,998) = 2.44, p = 0.119), although this difference was not significant.
- Female students disagreed with negative statements significantly more than male students (3.10 versus 2.94, F(1,995) = 4.965, p = 0.026). By disagreeing more with negative statements, they evaluated the tutors more positively.

Finally, we analyzed the responses on each statement one by one. In Table II, we have listed the normalized averages for male and female students on the 14 statements – note that female responses are almost always more positive than male responses. However, we did not find a significant main effect for sex on any of the statements.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.400</td>
<td>2.283</td>
</tr>
<tr>
<td>2</td>
<td>2.821</td>
<td>2.809</td>
</tr>
<tr>
<td>3</td>
<td>2.495</td>
<td>2.532</td>
</tr>
<tr>
<td>4</td>
<td>2.789</td>
<td>2.553</td>
</tr>
<tr>
<td>5</td>
<td>2.500</td>
<td>2.426</td>
</tr>
<tr>
<td>6</td>
<td>3.663</td>
<td>3.404</td>
</tr>
<tr>
<td>7</td>
<td>3.505</td>
<td>3.255</td>
</tr>
<tr>
<td>8</td>
<td>3.158</td>
<td>2.979</td>
</tr>
<tr>
<td>9</td>
<td>2.367</td>
<td>2.391</td>
</tr>
<tr>
<td>10</td>
<td>2.633</td>
<td>2.511</td>
</tr>
<tr>
<td>11</td>
<td>2.753</td>
<td>2.565</td>
</tr>
<tr>
<td>12</td>
<td>2.408</td>
<td>2.174</td>
</tr>
<tr>
<td>13</td>
<td>2.897</td>
<td>2.783</td>
</tr>
<tr>
<td>14</td>
<td>2.378</td>
<td>2.217</td>
</tr>
</tbody>
</table>

**II. for Loop Tutor Results**

Once again, we started by analyzing the difference between positively- and negatively-worded statements. We did a univariate ANOVA with feedback response as the dependent variable and sex and the type of feedback statement (positive versus negative) as fixed factors. Once again, we found a significant interaction between sex and the type of statement [F(1,1359) = 12.895, p = 0.000]:

- On learnability, we did not find a significant main effect for sex [F(1,707) = 1.124, p = 0.289].
- On user experience, we found a significant main effect for sex [F(1,568) = 4.946, p = 0.027]: female students rated their experience significantly more positively than male students (average of 3.053 versus 3.267).
- On usefulness, we did not find a significant main effect for sex [F(1,1286) = 0.142, p = 0.707].
- On usability, we did not find a significant main effect for sex [F(1,430) = 2.498, p = 0.115].
Female students agreed with positive statements significantly more than male students (average of 2.32 versus 2.46, \( F(1,680) = 4.879, p = 0.028 \));

Female students disagreed with negative statements significantly more than male students (3.22 versus 3.00, \( F(1,678) = 8.04, p = 0.005 \)).

Clearly, female students evaluated the tutor more positively than male students. When we normalized the data, i.e., reversed the Likert scale on negative statements and reanalyzed the feedback data, we found a significant main effect for sex \( F(1,1359) = 12.895, p = 0.000 \) – female students agreed significantly more with the feedback statements than male students (average of 2.55 versus 2.72).

Next, we analyzed the feedback by groups: the ability to learn from the tutor (statements 1–5), user experience (6–8,11), usefulness of the tutor (statements 9–10) and usability of the tutor (statements 12–14). We did a univariate ANOVA with normalized feedback response as the dependent variable and sex as the fixed factor.

- On learnability, we found a significant main effect for sex \( F(1,488) = 7.753, p = 0.006 \): female students agreed significantly more with the learnability statements than male students (average of 2.415 versus 2.626).

- On user experience, we did not find a significant main effect for sex \( F(1,191) = 2.731, p = 0.100 \).

- On usability, we did not find a significant main effect for sex \( F(1,191) = 2.731, p = 0.100 \).

- On the usefulness of the tutors, we found a marginally significant main effect for sex \( F(1,290) = 3.841, p = 0.051 \): female students agreed significantly more than male students (average of 2.368 versus 2.605).

Finally, we analyzed the responses on each statement one by one. We found a significant main effect for sex on the following two learnability statements:

- Statement 2 – The feedback provided to my answers was not clear: \( F(1,97) = 4.038, p = 0.047 \): female students disagreed significantly more than male students (reversed scale average 2.462 versus 2.831)

- Statement 4 – The feedback provided to my answers was not sufficient: \( F(1,97) = 5.662, p = 0.019 \): female students disagreed significantly more than male students (reversed scale average 2.462 versus 2.864)

In Table III, we have listed the normalized averages for male and female students on all 14 statements – note that female responses are almost always more positive than male responses.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.373</td>
<td>2.180</td>
</tr>
<tr>
<td>2</td>
<td>2.831</td>
<td>2.462</td>
</tr>
<tr>
<td>3</td>
<td>2.576</td>
<td>2.487</td>
</tr>
</tbody>
</table>

DISCUSSION

Female students not only disagreed more with negatively-worded statements than male students, but also agreed more with positively-worded statements than male students, and the difference was often significant. So, we conclude that female students evaluate our software tutors more positively than male students, and this is not an artifact of the positive versus negative wording of the feedback statements.

On learnability issues, we found a significant difference between male and female students on the for loop tutor: female students agreed significantly more with the learnability statements than male students (average of 2.415 versus 2.626). We tried to explain this difference based on who needed to use the tutor the most and who got to practice problem-solving with the tutor the most:

- We used pretest scores as a measure of who needed to use the tutor the most. We reasoned that students who needed to use the tutor the most might rate its learnability higher. But, the pretest scores of male and female students were nearly identical (0.605 for both groups, \( N=63 \) for male and \( N=45 \) for female students).

- We used the number of problems solved by students during practice as a measure of who got to practice with the tutor the most. We reasoned that students who practiced the most with the tutor might rate its learnability higher. We did not find a statistically significant difference between the number of practice problems solved by male (average 9.52, \( N=46 \)) and female students (average 10.87, \( N=31 \)) (\( p = 0.42 \)).

So, we could not explain the difference in learnability based on either the need to use the tutor or the amount of problem-solving practice received with the tutor.

When we analyzed the learnability statements one by one, as reported earlier, we found a significant difference between male and female students on both and only the two negatively-worded statements: statements 2 and 4. So, whereas female students assess software tutors more positively regardless of the wording of the statements on the feedback form, it may still be true that female students are less likely to agree with negatively-worded statements than male students.
On user experience, we found a significant difference between male and female students on the while loop tutor: female students rated their experience significantly more positively than male students (average of 3.053 versus 3.267). We tried to explain this difference based on how long students used the tutor and how rewarding an experience they had with the tutor:

- We used the number of problems solved by students as a measure of how long they used the tutor. We reasoned that students who solved more problems and hence, spent more time with the tutor would rate their user experience higher. We found that female students solved more problems during the tutoring session than male students: (average 19.89, N=54 versus average 16.56, N=112) and the difference was statistically significant ($p = 0.0233$). However, we did not find any correlation between the number of problems solved and the student’s feedback on user experience (Pearson coefficient 0.184).

- We used the percentage correctness of the students’ answers as a measure of how rewarding an experience they had with the tutor. We reasoned that the greater the percentage correctness of answers, the more rewarding the experience, and hence, the greater the likelihood of the student rating the user experience more favorably. But, the percentage correctness was not significantly different for male versus female students (0.681 for males, 0.709 for females, $p = 0.447$).

So, neither how long students used the tutor nor how rewarding an experience they had with the tutor explained the significant difference we found between male and female responses about user experience.

On usability issues, we did not find any significant difference between male and female students with either tutor. This might allude to an absence of gender bias in the user interface of the tutor.

Finally, on usefulness of the tutors, we found that female students were significantly more open to using tutors such as the for loop tutor than male students (average of 2.368 versus 2.605). We tried to explain this difference based on how rewarding an experience the students had with the tutor, measured in terms of the percentage correctness of their answers. We reasoned that when students found it rewarding to work with a tutor, they would be more likely to use such tutors again in the future. But, we found that male and female students had solved the same percentage of problems correctly (average 0.624 for males. N=64 and 0.626 for females, N=45).

In summary, we have not been able to explain the differences between male and female responses on learnability, user experience and usefulness of the tutors based on any intuitive criteria related to learning with the tutors. We also note that the differences are always in favor of female students rating the software tutors more positively than male students. We conclude that these differences are intrinsic to sex. In other words, female students assess our software tutors more positively than male students.

ACKNOWLEDGMENT

Partial support for this work was provided by the National Science Foundation’s Educational Innovation Program under grant CNS-0426021.

REFERENCES


