

Patterns in Student Assessment of Problem-Solving Software

Amruth N. Kumar

Ramapo College of New Jersey, amruth@ramapo.edu

Abstract - We investigated whether there were patterns in student assessment of (the usability, ability to learn from, and usefulness of) online problem-solving software tutors. We found that the correlation between the learning of students (how much they needed the software tutor, how long they worked with it, how much they learned from it) and their assessment of the tutor was weak. Further, we found that four-year college students assess the software tutor more favorably than two-year students, non-Caucasian students assess the tutor more favorably than Caucasian students, and female students assess the tutor more favorably than male students.

Index Terms – Assessment of tutor, Feedback, Learning, Problem-solving software.

INTRODUCTION

Software tutors to help students learn by solving problems are increasingly being developed and deployed in regular as well as blended courses (e.g., [1-3]). Typically, what a student thinks about the usability, usefulness and ability to learn from these tutors is subjective, i.e., subjective to the student, the specific tutor (design, user interface, clarity, etc.) and the experience the student has had using the tutor.

It would be interesting to find out whether there is any correlation between how much a student needs/learns using a software tutor and the student's opinions about the tutor; and whether there are any patterns of opinions that distinguish different demographic groups (status, race, sex, major, etc.). This information might help developers of software tutors better design and deploy the tutors. For instance, we have found that using software tutors helps improve the self-confidence of female students to be on par with that of male students [4], but, there is no correlation between the change in self-confidence and the change in learning of students using software tutors [5]. Female students assess software tutors more positively than male students [6] and this is not an artifact of the design of the feedback form [7].

In order to investigate whether there are any interesting patterns in the opinions of students about software tutors, we analyzed the data collected by a Computer Science software tutor on arithmetic expression evaluation over two years. In this paper, we will present the design and results of the study.

SOFTWARE TUTOR

For our study, we used a web-based software tutor on arithmetic expression evaluation. This tutor is part of a suite of tutors, called proplets (www.proplets.org) that we have developed to help *Computer Science I* students learn programming concepts. We chose the arithmetic expression evaluation tutor for our study because it is one of the first tutors used by students, and has the greatest number of students using it.

The arithmetic expression evaluation tutor presents expressions involving the correct evaluation, precedence and associativity of arithmetic operators used in programming languages such as C++, Java and C#. The student is asked to evaluate each expression one operator at a time. The tutor grades the student's answer and presents the correct evaluation of the expression as part of its feedback. The tutor presents problems on only the concepts the student does not yet understand, and presents problems on a concept until the student masters it. The tutoring session was limited to 30 minutes.

The software tutor was accessible over the web. When a student launched the tutor, the student went through the following stages:

- **Registration:** the student entered *demographic* information about himself/herself.
- **Tutoring:** the student solved problems covering over 20 concepts in arithmetic expression evaluation. Tutoring was structured as pre-test – practice – post-test so that the effect of using the tutor on the student's knowledge could be measured.
- **Feedback:** the student responded to 14 statements on the usability, usefulness and the ability to learn from the tutor. The student's responses were on a Likert scale and reflected the student's opinions about the software tutor.

DATA COLLECTION

The arithmetic expression evaluation tutor was used by students from 10 sections in spring 2007, and 15 sections in Fall 2007, Spring 2008 and Fall 2008. All the data was collected over the web, and combined for purposes of this study.

The feedback form contained the following 14 statements:

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 boring.
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.

Students were asked to respond to these statements on a Likert scale of 1 (Strongly Agree) to 5 (Strongly Disagree). Note that statements 1, 3, 5, 9, 11, 12 and 14 are positively-worded and the rest are negatively-worded. For purposes of this study, we re-coded the responses on negatively worded statements so that 1 always meant favorable opinion and 5 always meant unfavorable opinion. 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.

RESULTS

Quartile analysis: The score on the pre-test during the tutoring session is a measure of the prior preparedness of the student. We divided pre-test scores into performance quartiles and conducted ANOVA analysis to determine if those who fell into different quartiles provided different feedback responses. The following findings were all statistically significant (the sample size and averages for the quartiles are listed in Table 1 whenever the difference was statistically significant):

- The average of all feedback statements was rated less favorably by those who were in the bottom quartile than those who were in the top or second quartile. Additionally those who were in the top quartile rated these statements more favorably than those who were in the third quartile ($F(3, 993) = 9.84, p < .01, d = .03$).
- The average of all learning statements (1-5) was rated less favorably by those who were in the bottom quartile than those who were in the other three quartiles. Additionally those who were in the top quartile rated these statements more favorably than those who were in the third quartile ($F(3, 978) = 12.53, p < .01, d = .04$).
- The average of all usability statements (9-11) was rated more favorably by those who were in the top quartile than those who were in the bottom and third quartiles ($F(3, 879) = 5.36, p < .01, d = .02$).

TABLE 1:
SAMPLE SIZE AND AVERAGE OF THE QUARTILES FOR
AGGREGATE ANALYSIS OF FEEDBACK STATEMENTS

State ment	1st		2nd		3rd		4 th	
	N	Ave	N	Ave	N	Ave	N	Ave
1-14	252	2.21	265	2.31	242	2.41	238	2.49
1-5	251	2.30	263	2.42	238	2.50	230	2.65
9-11	227	2.02			206	2.22	210	2.26

When we did a post-hoc analysis of the data by individual statements, we found the same pattern on almost all the statements. The sample size and averages for the quartiles are listed in Table 2 whenever the difference was statistically significant.

- 1. “The generated problems were instructive” was rated less favorably by those who were in the bottom quartile than those who were in the other three quartiles ($F(3, 977) = 10.47, p < .01, d = .03$).
- 2. “The feedback provided to my answers was NOT clear” was rated less favorably by those who were in the bottom quartile than those who were in the other three quartiles ($F(3, 976) = 8.30, p < .01, d = .03$).
- 3. “The feedback provided to my answers was useful.” was rated less favorably by those who were in the bottom quartile than those who were in the top quartile ($F(3, 976) = 3.05, p < .05, d = .01$).
- 4. “The feedback provided to my answers was NOT sufficient” was rated less favorably by those who were in the bottom quartile than those who were in the other three quartiles ($F(3, 974) = 9.86, p < .01, d = .03$).
- 6. “Using this tutor to learn was time-consuming” was rated less favorably by those who were in the bottom quartile than those who were in the top or second quartile. Additionally those who were in the top quartile rated this statement more favorably than those who were in the third quartile ($F(3, 972) = 10.63, p < .01, d = .03$).
- 7. “The generated problems were repetitive and boring” was rated less favorably by those who were in the bottom quartile than those who were in the top or second quartile. Additionally those who were in the top quartile rated this statement more favorably than those who were in the third quartile ($F(3, 974) = 6.42, p < .01, d = .02$).
- 8. “The progress of my learning was NOT presented clearly” was rated less favorably by those who were in the bottom quartile than those who were in the other three quartiles ($F(3, 971) = 6.20, p < .01, d = .02$).
- 10. “It was NOT easy to learn how to use this tutor” was rated less favorably by those who were in the bottom quartile than those who were in the top or second quartile ($F(3, 876) = 6.74, p < .01, d = .02$).
- 14. “I would like to see such tutors on other topics” was rated less favorably by those who were in the third quartile than those who were in the top quartile ($F(3, 873) = 3.48, p < .05, d = .01$).

TABLE 2:
SAMPLE SIZE AND AVERAGE OF THE QUARTILES FOR
INDIVIDUAL ANALYSIS OF FEEDBACK STATEMENTS

Statement	1 st		2nd		3rd		4th	
	N	Ave	N	Ave	N	Ave	N	Ave
1	251	1.80	262	1.87	238	1.90	230	2.17
2	250	2.17	263	2.35	237	2.38	230	2.67
3	250	1.92					230	2.12
4	250	2.16	263	2.28	237	2.35	228	2.64
6	250	2.70	223	2.91	235	3.09	229	3.24
7	250	2.80	263	2.95	237	3.08	228	3.20
8	249	2.51	262	2.73	236	2.75	228	2.90
10	226	1.99	240	2.07			209	2.38
14	225	2.04			204	2.27		

Clearly, the bottom quartile was less favorable in its rating of the usability, usefulness and ability to learn from our tutor than the other three quartiles.

2-year versus 4-year institutions: We used *t*-test to compare opinions of students from 2-year versus 4-year institutions. Those in 4-year (baccalaureate) institutions rated the following more favorably than those in 2-year (community) colleges (sample size and averages listed in Table 3):

- 2. “The feedback provided to my answers was NOT clear” ($t(980) = 2.61, p < .01, d = .17$).
- 4. “The feedback provided to my answers was NOT sufficient” ($t(978) = 2.50, p < .01, d = .16$).
- 7. “The generated problems were repetitive and boring” ($t(978) = 2.39, p < .05, d = .15$).
- 10. “It was NOT easy to learn how to use this tutor” ($t(880) = 2.66, p < .01, d = .18$).

TABLE 3:
SAMPLE SIZE AND AVERAGE OF 4-YEAR AND 2-YEAR
STUDENTS

Statement	4-year		2-year	
	N	Ave	N	Ave
2	653	2.32	329	2.52
4	653	2.30	327	2.46
7	652	3.06	328	2.89
10	584	2.09	298	2.28

This could suggest that students in 2-year colleges are less favorable in their assessment of the software tutor. Since all four statements in question (2,4,7 and 10) are negatively worded, we cannot rule out some bias against agreeing with negatively worded statements among students from 4-year institutions.

Race: We compared the responses of Caucasian and non-Caucasian students using *t*-test, and found the following significant differences (sample size and averages in Table 4):

- The average of all feedback statements was rated more favorably by non-Caucasians than Caucasians ($t(878) = 2.16, p < .05, d = .16$).
- The average of all statements related to usefulness was rated more favorably by non-Caucasians than Caucasians ($t(773) = 3.51, p < .01, d = .25$).

TABLE 4:
SAMPLE SIZE AND AVERAGE OF CAUCASIAN AND NON-
CAUCASIAN STUDENTS – AGGREGATE ANALYSIS

Statement	Non-Caucasians		Caucasians	
	N	Ave	N	Ave
1-14	322	2.38	558	2.89
12-14	277	2.10	498	2.31

When we did a post-hoc analysis of the data by individual statements, we found the same pattern on almost all the statements. The following statements were more favorably rated by non-Caucasian than Caucasian students (sample size and averages are listed in Table 5):

- 3. “The feedback provided to my answers was useful” ($t(865) = 3.45, p < .01, d = .23$).
- 7. “The generated problems were repetitive and boring” ($t(863) = 2.02, p < .05, d = .14$).
- 11. “It was clear to me after each problem, how much I knew and how much I had yet to learn.” ($t(772) = 5.01, p < .01, d = .17$).
- 12. “This tutor should be made available to all the students.” ($t(773) = 3.67, p < .01, d = .27$).
- 14. “I would like to see such tutors on other topics” ($t(772) = 3.76, p < .01, d = .27$).

One exception was statement 6. “Using this tutor to learn was time-consuming”, which was rated less favorably by non-Caucasians (N = 314, Ave = 3.08) than Caucasians (N = 550, Ave = 2.92) ($t(862) = -2.03, p < .05, d = .14$).

TABLE 5:
SAMPLE SIZE AND AVERAGE OF CAUCASIAN AND NON-
CAUCASIAN STUDENTS – INDIVIDUAL ANALYSIS

Statement	Non-Caucasians		Caucasians	
	N	Ave	N	Ave
3	314	1.90	551	2.40
7	313	2.90	552	3.05
11	277	2.06	497	2.40
12	277	1.90	498	2.13
14	277	1.98	497	2.23

So, more often than not, non-Caucasian students rated the tutor more favorably than Caucasian students.

Male versus Female students: We used *t*-test to compare the responses of male and female students. The following findings were all statistically significant (the sample size and averages for male and female students are listed in Table 6):

- The aggregate average of females students on all 14 feedback statements was more favorable than that of male students ($t(982) = -2.79, p < .01, d = .18$).
- Females rated learning statements (1-5) more favorably than males ($t(967) = 2.85, p < .01, d = .18$).
- Females rated usability statements (9-11) more favorably than males ($t(871) = -2.22, p < .05, d = .15$).
- Females rated usefulness statements (12-14) more favorably than males ($t(866) = -3.02, p < .01, d = .21$).

TABLE 6:
SAMPLE SIZE AND AVERAGE OF FEMALE AND MALE STUDENTS
- AGGREGATE ANALYSIS

Statement	Female		Male	
	N	Ave	N	Ave
1-14	279	2.27	705	2.39
1-5	278	2.37	691	2.50
9-11	247	2.05	626	2.17
12-14	246	2.12	622	2.30

When we did a post-hoc analysis of the data by individual statements, we found the same pattern on almost all the statements - All the responses that were statistically significantly different between males and females were rated more favorably by females. Females rated the following statements more favorably than males (sample size and averages are listed in Table 7):

- 4. “The feedback provided to my answers was not sufficient” ($t(963) = -2.07, p < .05, d = .13$).
- 7. “The generated problems were repetitive and boring” ($t(963) = -2.96, p < .01, d = .19$).
- 8. “The progress of my learning was NOT presented clearly” ($t(960) = -2.98, p < .01, d = .19$).
- 9. “It was easy to use this tutor” ($t(871) = -2.26, p < .05, d = .15$).
- 10. “It was not easy to learn how to use this tutor” ($t(868) = -2.36, p < .05, d = .16$).
- 13. “If this tutor is made available, I would NOT use it” ($t(864) = -2.93, p < .01, d = .20$).
- 14. “I would like to see such tutors on other topics” ($t(865) = -2.75, p < .01, d = .19$).

TABLE 7:
SAMPLE SIZE AND AVERAGE OF FEMALE AND MALE STUDENTS
- INDIVIDUAL ANALYSIS

Statement	Female		Male	
	N	Ave	N	Ave
4	278	2.25	687	2.40
7	227	2.84	688	3.07
8	277	2.57	685	2.79
9	247	2.03	623	2.21
10	247	2.03	623	2.21
13	244	2.36	622	2.59
14	246	2.02	621	2.21

These results are intriguing. They reinforce the results we had obtained from the analysis of the data collected by two loop tutors in spring 2007 [7] and multiple tutors in fall 2004 and spring 2005 [6]. Given that we have been repeatedly seeing this pattern in multiple sets of data, the pattern is not an anomaly or artifact of a specific tutor or semester.

Correlation between learning and feedback:

- Pre-test scores are indicative of the prior preparation of students. We found a statistically significant, but weak correlation between pre-test scores and response on feedback statements, as shown in Table 8, Column 2 (Table 8 lists only statistically significant correlations). As discussed under quartile analysis, all these

correlations were negative, i.e., the greater the pre-test score, the more positive the feedback response.

- We found statistically significant, but weak positive correlation between the total time spent on tutoring and response on feedback statements, i.e., the more time spent, the less positive the feedback response (Table 8, Column 3).
- Practice score is indicative of how much the student learned from the tutoring session. We found statistically significant, but weak negative correlation between practice score and response on feedback statements, i.e., the more the student learned, the more positive the feedback response (Table 8, Column 4).

TABLE 8:
STATISTICALLY SIGNIFICANT CORRELATION OF FEEDBACK
RESPONSES WITH PRE-TEST SCORE, TOTAL TIME SPENT ON
TUTORING, AND PRACTICE SCORE

Statement	Pre-test Score	Total Time	Practice Score
1-14	-0.18	0.12	-0.19
1	-0.17	0.12	-0.12
2	-0.15	0.11	-0.20
3	-0.10		-0.20
4	-0.18	0.16	-0.20
6	-0.18	0.20	-0.17
7	-0.13		-0.14
8	-0.12	0.11	
9		0.13	-0.12
10	-0.18	0.16	
12			-0.10
13			-0.12
14	-0.10		

Although we found several statistically significant correlations between feedback responses and pre-test score, total time spent on tutoring, and practice score, all the correlations were weak. All the correlations between feedback responses and pre-test score, and practice score were negative, reinforcing the results from quartile analysis.

DISCUSSION

The following are some of the patterns we have observed in feedback responses:

- The students with the least prior preparation among those who used the tutor provided the least favorable responses on the usability, usefulness and ability to learn from the tutors. This could be due to disengagement on the part of the least-prepared students, and might call for more motivational messages to be provided by the tutor in order to encourage them to engage with the tutor.
- Students in 4-year institutions rated the tutor more favorably than those in 2-year institutions. It is not yet clear if this difference is purely an artifact of the negative wording of the four statements on which the difference was observed, and needs to be investigated further.

ACKNOWLEDGMENT

Partial support for this work was provided by the National Science Foundation under grant DUE-0817187.

REFERENCES

- [1] Castro-Schez, J.J., et al., *Designing and Using Software Tools for Educational Purposes: FLAT, A Case Study*. IEEE Transactions on Education, 2009. **52**(1): p. 66-74.
- [2] Weyten, L., P. Rombouts, and J.D. Maeyer, *Web-Based Trainer for Electrical Circuit Analysis*. IEEE Transactions on Education, 2009. **52**(1): p. 185-189.
- [3] Nestor, J.A., *Experience with the CADAPPLETS Project*. IEEE Transactions on Education, 2008. **51**(3): p. 342-348.
- [4] Kumar, A.N. *The Effect of Using Problem-Solving Software Tutors on the Self-Confidence of Female Students*. in *39th SIGCSE Technical Symposium*. 2008. Portland, OR: p 523-527.
- [5] Kumar, A.N. *The Effect of Using Problem-Solving Tutors on the Self-Confidence of Students*. in *18th Annual Psychology of Programming Workshop (PPIG 06)*. 2006. Brighton, U.K: p 275-283.
- [6] Kumar, A.N. *Do female students feel differently than male students about using software tutors?* in *Frontiers in Education Conference (FIE 2006)*. 2006. San Diego, CA.
- [7] Kumar, A.N. *Female Students Assess Software Tutors More Positively Than Male Students*. in *Frontiers in Education Conference (FIE 2008)*. 2008. Saratoga Springs, NY.

- Non-Caucasian students rated the tutor more favorably than Caucasian students. This could be a result of cultural differences, and needs to be investigated further.
- Female students consistently rated the tutor more favorably than male students. We have found this pattern in earlier evaluations and reported it [6,7]. We plan to look into gender issues that might explain this consistent difference between male and female students.
- Correlations between feedback responses and prior-preparedness, time spent with the tutor, and learning affected by the tutor – are all weak, even when statistically significant. The more the prior-preparedness or learning affected by the tutor, the more favorable the rating of the tutor by the student.

Effect sizes (d value) are practically significant for the differences between 4-year and 2-year students, non-Caucasian and Caucasian students, and male and female students. So, we plan to investigate these differences further in future evaluations of our tutors.

We plan to repeat this study with data collected by other types of software tutors that we have developed (tutors on debugging programs and tutors on predicting the output of programs) to see whether or not these patterns are specific to the expression evaluation tutor, or can be found across different types of tutors.