

# Results from Repeated Evaluation of an Online Tutor on Introductory Computer Science

Amruth N. Kumar

Ramapo College of New Jersey, amruth@ramapo.edu

**Abstract** – We analyzed the data collected over 7 semesters by a single Computer Science software tutor to study the differences between the sexes and races on their prior self-confidence, prior preparedness and their assessment of the tutor. We found that when there was a statistically significant difference in the prior self-confidence of male and female students, female students had lower prior self-confidence than male students, in spite of the fact that there was no significant difference in the prior preparedness of male and female students. The prior self-confidence of female students in Computer Science may be improving with increasing enrollment. Whenever there was a statistically significant difference among racial groups, positively stereotyped racial groups were better prepared and had higher prior self-confidence than the traditionally under-represented racial groups. Whenever there was a statistically significant difference between the sexes in the assessment of the tutor, female students assessed the tutor more favorably than male students. When there was a statistically significant difference between racial groups, under-represented racial groups assessed the tutor more favorably than positively stereotyped racial groups. When there was a statistically significant difference in how developer’s students assessed the tutor versus how other adopters’ students assessed it, assessment by developer’s students was more positive than that by students of other adopters.

*Index Terms* – Computer Science, Feedback, Online Tutor, Race, Self-confidence, Sex.

## INTRODUCTION

We analyzed the data collected over 7 semesters by a Computer Science software tutor on arithmetic expression evaluation. We studied the differences between the sexes and races on their prior self-confidence, prior preparedness, and their assessment of the tutor.

An advantage of using data from 7 semesters of the same tutor for the study is that it eliminates the effect of inevitable variations in student characteristics from one semester to the next. Another is that finding the same pattern in the data from multiple semesters adds greater validity to the results.

Some of the hypotheses we considered in our study were re-evaluations of earlier results, although using data from one tutor in multiple semesters rather than multiple

tutors in one semester/year. The re-evaluations helped us qualify earlier results. Other hypotheses that we considered were new, and could be tested because of the availability of data for a larger sample size of students, who had all used the same exact experimental setup.

We chose the tutor on arithmetic expression evaluation because it is the first tutor used in introductory Computer Science courses. Therefore, it is used by the most number of students, and they are unaffected by the use of any earlier software tutors in the course.

## THE TUTOR AND EXPERIMENTAL SETUP

The tutor on arithmetic expressions presents problems on correct evaluation of arithmetic operators: addition, subtraction, multiplication, division and remainder as applied to whole and real operands. It also covers potential errors in arithmetic expressions such as dividing by zero, and inapplicability of remainder operator to integer operands in C++. It is available for C, C++, Java and C# programming languages. In all, the tutor covers 25 learning objectives such as correct evaluation, precedence, and associativity of the various arithmetic operators.

When students use the arithmetic expression tutor, they work through pre-survey, tutoring, feedback, and post-survey stages back to back, all administered online.

During pre-survey, students fill out a self-confidence survey consisting of five statements, to which they respond on a five-point Likert scale of Very well, Well, Average, Not well, and Not at all. The statements are:

1. How well do you know Arithmetic Operators (+, -, \*, /, %)?
2. How well do you know integer division?
3. How well do you know the Remainder Operator (%)?
4. How well do you know operator precedence concepts?
5. How well do you know operator associativity concepts?

During post-survey, students respond to the same five statements, although, after the tutoring stage. Their responses on the pre-survey are a measure of their self-confidence in their knowledge of arithmetic expressions before using the tutor. The difference between pre-survey and post-survey scores is a measure of the change in self-confidence due to their use of the tutor.

During the tutoring stage, students work through pre-test-practice-post-test protocol. The total time allowed for the tutoring stage is 30 minutes.

1. During pretest, students solve 16 problems that cover the 25 learning objectives. Their pre-test score is a

measure of how much they already knew the material before using the tutor.

2. During practice, students are presented additional practice problems on only those learning objectives on which they had solved problems incorrectly during pre-test. If a student solved all the pre-test problems correctly, the student is not presented any practice problems. If on the other hand, the student solved all the problems incorrectly during pre-test, the number of problems presented to the student for practice is only limited by the time allocated for the tutoring stage.
3. During post-test, students are presented problems on only those learning objectives on which they got sufficient practice.

During the feedback stage, students fill out a survey consisting of 14 statements on the usability, learnability and usefulness of the tutor. They respond to each statement on a five-point Likert-scale of Strongly agree, Agree, Neutral, Disagree and Strongly disagree. The 14 statements are:

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.

Note that positively and negatively-worded statements alternate in the feedback form. Their responses on the feedback form were used as a measure of their assessment of the tutor.

**DATA COLLECTION AND ANALYSIS**

For the purposes of this study, data of arithmetic tutor was used from 7 semesters: spring 2007 through spring 2010. The same pre-survey, post-survey, and feedback instruments were administered during all 7 semesters. The same pre-test was also used during all 7 semesters. Each semester, the tutor was used in a controlled experiment to test a different hypothesis/treatment, such as the effectiveness of providing error-detection, but not error-correction support during practice. Therefore, data from practice and post-test was not used for this study.

Pre-survey, post-survey and feedback instruments were coded so that 1 was the most positive response and 5 the most negative response. For analysis purposes, responses on negatively-worded feedback statements were reversed. In

other words, the lower the score, the more positive the response, whether on pre-survey, post-survey or feedback.

On pre-test, each of the 16 problems was worth 1 point. For analysis purposes, students who solved 10 or more problems were considered. Since not all the students solved all 16 pre-test problems in the allowed time, score per problem (range: 0 → 1.0) rather than raw score (range: 0 → 16.0) was used for analysis.

Table 1 lists the number of students who used the tutor in each of the 7 semesters, as well as the number of male and female students. (Participants were asked to identify their sex (biological notion of male/female) rather than their gender (social/cultural notion of man/woman) [14].) Table II lists participants by the type of institution, the host institution itself being a 4-year institution. Table III lists participants by race. Since some students did not indicate their sex/race, the sum of all the races/sexes may not add up to the total number of students each semester.

TABLE I  
PARTICIPANTS BY SEX

Semester	N	Male	Female
Spring 2007	255	188	65
Fall 2007	258	161	94
Spring 2008	280	194	73
Fall 2008	255	194	56
Spring 2009	452	315	125
Fall 2009	295	187	87
Spring 2010	343	250	83

TABLE II  
PARTICIPANTS BY INSTITUTION

Semester	Host	Other 4-year	2-year
Spring 2007	45	92	118
Fall 2007	71	130	57
Spring 2008	51	114	114
Fall 2008	55	142	58
Spring 2009	33	365	53
Fall 2009	52	201	42
Spring 2010	43	242	58

TABLE III  
PARTICIPANTS BY RACE

Semester	Cauc.	Asian	Black	Hisp.	Native
Spring 2007	141	56	10	14	2
Fall 2007	137	23	10	34	13
Spring 2008	149	34	25	18	2
Fall 2008	147	37	13	7	5
Spring 2009	273	73	25	19	7
Fall 2009	156	39	13	13	1
Spring 2010	218	42	23	20	2

**PRIOR SELF CONFIDENCE AND PRE-TEST SCORE**

Literature review reveals that historically, female students have had lower self-confidence than male students in

computer-related abilities [1, 2]. Women are less confident than men in their ability to achieve their educational goals in computing at the undergraduate [3] and the graduate level [4]. Women have a tendency to enter computing classes with considerably less confidence than men [5]. In an earlier study conducted using 5 tutors in spring 2006 [6], we had found this to be true on 3 of the tutors, not including the arithmetic expressions tutor.

In the current study, independent samples t-test showed significant difference in the cumulative pre-survey responses of male and female students [ $t(1014) = -2.537, p = 0.011$ ]: 12.27 for male (N=1342) versus 12.83 for female students (N=538). Since the lower the score, the more self-confident the response, male students had greater self-confidence before using the tutor than female students. ANOVA analysis yielded significant main effect for sex [ $F(1,1879) = 5.516, p = 0.019$ ] as well as semester [ $F(6,1879) = 2.938, p = 0.007$ ], but no significant interaction between sex and semester. Figure 1 shows the cumulative pre-survey responses of male and female students through the 7 semesters. However, the difference between the sexes was statistically significant in only 3 of the semesters:

- Fall 2008 [ $t(76.45) = -1.631, p = 0.107$ ] (marginally significant, assuming unequal variances): 12.72 (N=177) for male versus 13.87 for female students (N=47)
- Spring 2009 [ $t(217.40) = -2.759, p = 0.006$ ]: 12.24 for male (N=290) versus 13.48 for female students (N=121)
- Fall 2009 [ $t(181.70) = -1.85, p = 0.066$ ]: 12.15 for male (N=175) versus 13.11 for female students (N=81)

In all three cases, cumulative responses of female students were higher than, and hence, their prior self-confidence was lower than that of male students.

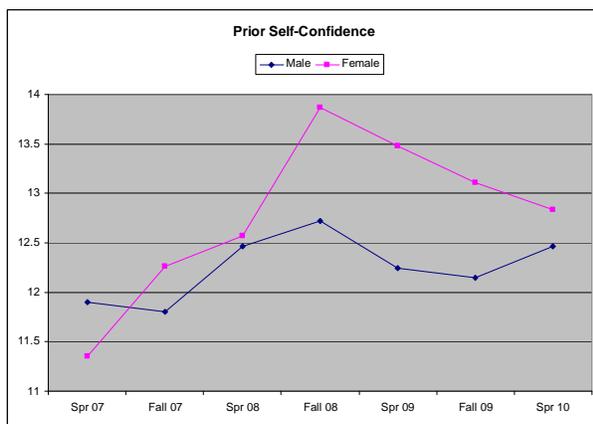


FIGURE 1  
CHANGE IN THE PRIOR SELF-CONFIDENCE OF STUDENTS OVER THE 7 SEMESTERS: MALE VS FEMALE.

In the earlier study [6], we had reported that the self-confidence of female students before using the software tutors was in many cases lower than that of male students. As a result of the current study, we qualify the earlier result as follows: *when there was a statistically significant difference in the prior self-confidence of male and female*

students, female students had lower prior self-confidence than male students.

Ironically, female students have lower prior self-confidence even when they have the same level of skills as male students [7]. This is borne out by analysis of the pre-test score of students on the tutor: ANOVA analysis yielded no significant main effect for sex [ $F(1,2070) = 0.651, p = 0.42$ ], i.e., there was no significant difference in the pre-test scores of male and female students as shown in Figure 2, notwithstanding the lower prior self-confidence of female students.

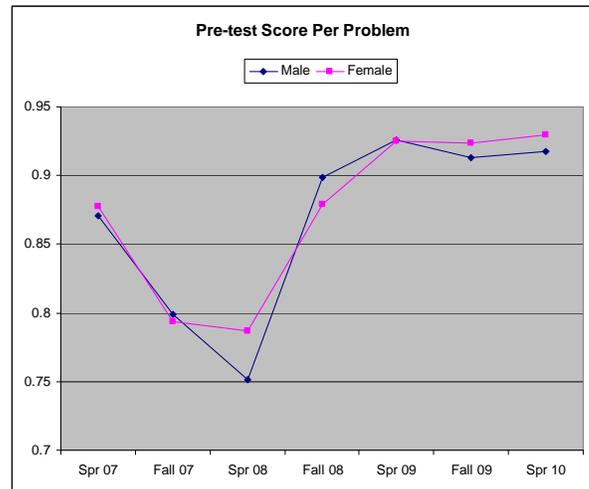


FIGURE 2  
CHANGE IN THE SCORE PER PROBLEM ON PRE-TEST OVER THE 7 SEMESTERS: MALE VS FEMALE.

ANOVA analysis showed that the change in the prior self-confidence of female students over the 7 semesters was statistically significant [ $F(6,537) = 2.576, p = 0.018$ ]. The same was not true for male students [ $F(6,1341) = 0.936, p = 0.468$ ]. Given this, the shape of the change as seen in Figure 1 is encouraging: cumulative survey response of female students monotonically increased until fall 2008, and has been decreasing since then. In other words, prior self-confidence of female students monotonically decreased till fall 2008 and has been increasing ever since. The shape shown in Figure 1 is roughly the inverse of enrollment trend in Computer Science – according to the latest available Taulbee survey (2008-09, [www.cra.org](http://www.cra.org)), the number of newly declared undergraduate Computer Science/Computer Engineering majors decreased from 2002 through 2007 and has been gradually increasing since then. Research shows that gender gap in computer self-efficacy is closing [8], which could also explain the increasing prior self-confidence of female students.

Next, we considered the prior self-confidence of Caucasians versus minorities. Independent samples t-test showed no significant difference in the cumulative pre-survey responses of Caucasians and non-Caucasians (including Asian, African, Hispanic and Native Americans) [ $t(973.76) = 0.676, p = 0.499$ ]. Asians are positively stereotyped in quantitative domains such as Computer

Science (e.g., [9]). Therefore, we repeated the analysis grouping Caucasians and Asians together, i.e., we compared the racial groups that are positively stereotyped in Computer Science against the minorities that are traditionally considered under-represented in Computer Science [10]. Independent samples t-test yielded a significant difference between Caucasians + Asians and the under-represented races [ $t(333.90) = 2.491, p = 0.013$ ]: cumulative pre-survey response of Caucasians + Asians was 12.28 (N=1416) versus 13.03 for the other races (N=246).

ANOVA analysis yielded significant main effect for positive stereotype (Caucasians + Asians versus Others) [ $F(1,1661) = 8.485, p = 0.004$ ]. Figure 3 shows the cumulative pre-survey responses of Caucasians + Asians and other students through the 7 semesters. The difference between the two groups was statistically significant in two semesters:

- Fall 07 [ $t(188) = -3.242, p = 0.001$ ], with 11.47 for Caucasians + Asians (N=143) and 13.94 for other racial groups (N=47).
- Fall 09 [ $t(26.14) = -2.020, p = 0.054$ ], with 12.32 for Caucasians + Asians (N=190) and 14.35 for other racial groups (N=23).

Note that in both cases, the positively stereotyped racial groups had higher prior self-confidence.

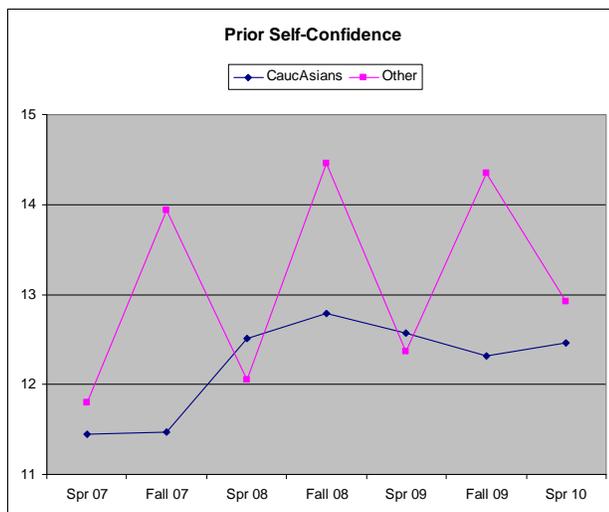


FIGURE 3

CHANGE IN PRIOR SELF-CONFIDENCE OVER THE 7 SEMESTERS: POSITIVELY STEREOTYPED VS UNDER-REPRESENTED RACIAL GROUPS

ANOVA analysis of pre-test score also yielded significant main effect for positive stereotype [ $F(1,1800) = 25.17, p < 0.001$ ] – Caucasians + Asians scored an average of 0.89 (out of 1.0) whereas students in under-represented racial groups scored an average of 0.825. Figure 4 shows the change in pre-test score of the two groups over the 7 semesters. The difference between the two groups was statistically significant in four of the 7 semesters:

- Fall 07 [ $t(215) = 4.983, p < 0.001$ ]: 0.828 for Caucasians + Asians (N=160) versus 0.708 for other racial groups (N=57).

- Fall 08 [ $t(207) = 1.685, p = 0.093$ ] (marginally significant): 0.906 for Caucasians + Asians (N=184) versus 0.863 for other racial groups (N=25).
- Spring 09 [ $t(395) = 3.157, p = 0.002$ ]: 0.932 for Caucasians + Asians (N=346) versus 0.878 for other racial groups (N=51).
- Fall 09 [ $t(220) = 3.938, p < 0.001$ ]: 0.935 for Caucasians + Asians (N=195) versus 0.858 for other racial groups (N=27).

Note that in all four cases, Caucasians + Asians scored higher than the under-represented racial groups. We conclude that *whenever there was statistically significant difference among racial groups, positively stereotyped racial groups (Caucasians and Asians) were better prepared and had higher prior self-confidence than the under-represented racial groups.*

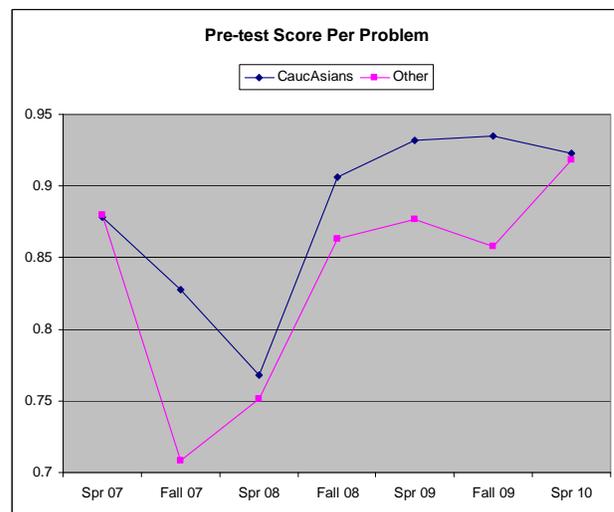


FIGURE 4

CHANGE IN THE SCORE PER PROBLEM ON PRE-TEST OVER THE 7 SEMESTERS: POSITIVELY STEREOTYPED VS UNDER-REPRESENTED RACIAL GROUPS.

### ASSESSMENT OF THE TUTOR

Students assessed the usability, learnability and usefulness of the tutor during the feedback stage. We considered whether there was an intrinsic difference between the sexes or the races in how they assessed the tutor. Since feedback stage followed tutoring, where control and test groups got different treatments, we included treatment as a fixed factor during data analysis.

In a study conducted in fall 2004 and spring 2005 using four different tutors [11], we had found that there was no statistically significant difference between male and female students, except on two of the 14 feedback statements. At that time, these were the only two negatively-worded statements on the feedback form. So, the difference could have been purely an artifact of the negative wording of the two statements. Therefore, we re-worded and reordered the feedback form to contain equal number of positively and negatively-worded statements, and conducted a follow-up study using two loop tutors in spring 2007 [12]. Once again,

we found that female students assessed the two tutors more positively than male students, and this was not merely an artifact of positive versus negative wording of the feedback statements, i.e., female students both agreed with positively worded statements and disagreed with negatively worded statements more than male students.

In the current study, we considered whether we could find the same difference between the sexes over multiple semesters. Independent samples t-test showed a significant difference in the cumulative feedback responses of male and female students [t(1673) = 4.494, p < 0.001]: 32.17 for males (N=1203) versus 30.14 for females (N=472). Since the lower the score, the more positive the feedback, female students assessed the tutor more positively than male students. ANOVA analysis yielded significant main effect for sex [F(1,1674) = 18.622, p < 0.001] as well as semester [F(6,1674) = 5.242, p < 0.001], but no significant main effect for treatment, nor any significant interaction between sex and semester, or sex and treatment.

Figure 5 shows the cumulative feedback responses of male and female students through the 7 semesters. The difference was statistically significant in four of the semesters:

- Spring 2007 [t(94.8) = 2.433, p = 0.017]: 33.61 for male (N=153) versus 30.76 for female students (N=54).
- Fall 2007 [t(208) = 2.278, p = 0.024]: 33.39 for male (N=133) versus 30.61 for female students (N=77).
- Fall 2009 [t(221) = 2.111, p = 0.036]: 30.33 for male (N=151) versus 27.94 for female students (N=72).
- Spring 2010 [t(282) = 2.585, p = 0.01]: 31.86 for male (N=213) versus 28.92 for female students (N=71).

In all four semesters, female students assessed the tutor more favorably than male students. As a result of the current study, we qualify the earlier result as follows: *when there was a statistically significant difference between male and female students' assessment of the tutor, female students assessed the tutor more favorably than male students.*

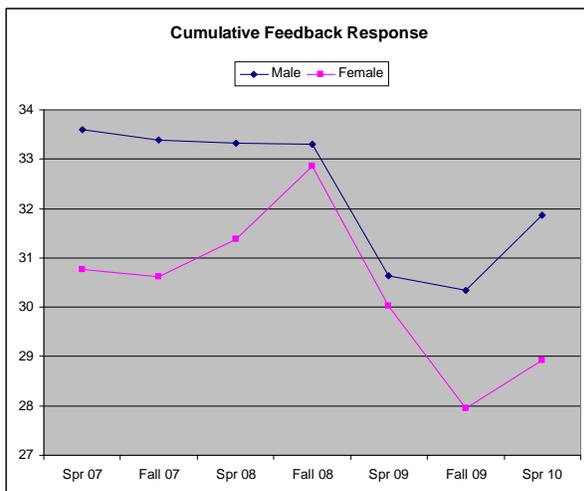


FIGURE 5

CHANGE IN THE ASSESSMENT OF STUDENTS OVER THE 7 SEMESTERS: MALE VS FEMALE.

Next, we considered the difference in assessment by positively stereotyped racial groups (Caucasians + Asians) versus under-represented racial groups. Independent samples t-test of cumulative feedback response showed a marginally significant difference between Caucasians + Asians versus students from under-represented (Black, Hispanic and Native American) racial groups [t(285.62) = 1.871, p = 0.062]: the cumulative response of Caucasians + Asians was 31.54 (N=1263) as compared to 30.40 for other racial groups (N=210), i.e., under-represented racial groups assessed the tutor *more* favorably than Caucasians + Asians.

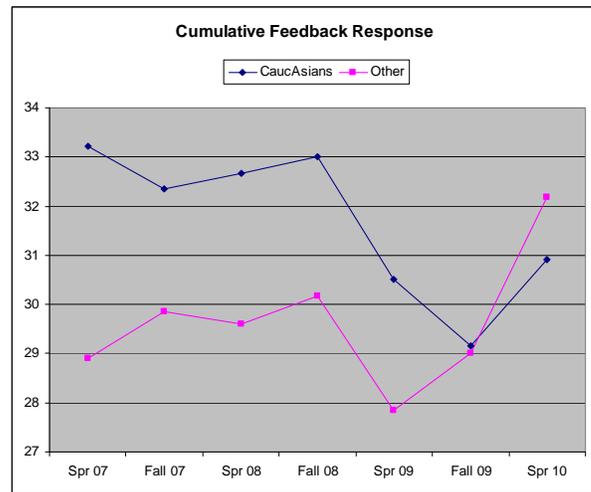


FIGURE 6

CHANGE IN THE ASSESSMENT OF STUDENTS OVER THE 7 SEMESTERS: POSITIVELY STEREOTYPED VS UNDER-REPRESENTED RACIAL GROUPS.

ANOVA analysis of cumulative feedback response yielded a significant main effect for positive stereotype [F(1,1472) = 7.541, p = 0.006], but not treatment. Figure 6 shows change in the cumulative feedback response of the two racial groups over the 7 semesters. The difference between the two groups was marginally significant in three of the 7 semesters:

- Spring 07 [t(28.12) = 1.775, p = 0.087]: 33.34 for Caucasians + Asians (N=163) versus 30.45 for other racial groups (N=22).
- Spring 08 [t(41.7) = 1.944, p = 0.059]: 33.08 for Caucasians + Asians (N=152) versus 29.79 for other racial groups (N=29).
- Spring 09 [t(52.62) = 1.96, p = 0.055]: 30.35 for Caucasians + Asians (N=277) versus 27.98 for other racial groups (N=40).

We conclude that *whenever there was a statistically significant difference between racial groups in their assessment of the tutor, under-represented racial groups assessed the tutor more favorably than positively stereotyped groups (Caucasians and Asians).* This result clarifies the preliminary result we had reported earlier that non-Caucasian students assessed tutors more favorably than Caucasian students [13]. One explanation for this could be that because under-represented racial groups were less well-prepared before using the tutor, they stood to benefit more

from the tutor, and therefore, assessed the tutor more positively. But, previously, we had found that correlation between student learning (how much they needed a tutor, how long they worked with it, and how much they learned from it) and their assessment of the tutor was weak [13].

We studied whether there was any difference in the assessment of the tutor between students of the developer of the tutor and students of adopters not related to the developer or his institution. For this analysis, we considered only 4-year institutions since the developer is at a 4-year institution. We considered only those institutions that were in the same treatment group each semester as students of the developer, e.g., since all the developer's students were in control group in spring 2007, they were compared against only the students from other 4-year institutions who were also in control group. Statistically significant difference was observed between the developer's students and students from other institutions in four of the seven semesters:

- [t(20.48) = -2.413, p = 0.025] in fall 07 test group: 24.47 for developer's students (N=15) versus 30.45 for others (N=31).
- [t(31.72) = -2.313, p = 0.027] in spring 08 control group: 27.50 for developer's students (N=16) versus 32.80 for others (N=25).
- [t(12.59) = -2.912, p = 0.012] in spring 09 control group: 26.09 for developer's students (N=11) versus 30.97 for others (N=217).
- [t(133) = -2.077, p = 0.04] for spring 10 control group: 26.67 for developer's students (N=18) versus 30.90 for others (N=117).

Remarkably, every time there was statistically significant difference between the two groups, the average response of developer's students was less than that of students of adopters at other 4-year institutions. We conclude that *whenever there was a statistically significant difference in how developer's students assessed a tutor versus how other adopters' students assessed it, the assessment by developer's students was more positive than that by students of other adopters.*

The reasons for positive assessment by developer's students are not clear. Students might assess a tutor more positively if they perceive an institutional connection to it. But, the developer does not introduce the tutor in class as "locally developed". It is quite possible that adopters at other institutions refer to the tutor in class as "externally developed". The address at which the tutor is accessed on the web would hint at the provenance of the tutor, but it is not necessarily true that students had to individually enter the web address. An alternative explanation is that the teaching style of the developer is more congruent with the types of problems presented by the tutor. This hypothesis warrants further research.

As explained in the introduction, arithmetic expressions tutor is the first tutor used by students in the introductory Computer Science course. In all, 14 such software tutors

have been developed for the introductory Computer Science course ([www.problets.org](http://www.problets.org)), and are being used each semester by multiple institutions. As part of future work, we plan to repeat this study with a more advanced tutor to see if the results are dependent on the level of complexity of the tutor topic.

### ACKNOWLEDGMENT

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

### REFERENCES

- [1] Beckwith, L. and Burnett, M. Gender: An Important Factor in Problem-Solving Software? Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing Languages and Environments. IEEE Press. 2004. 107-114
- [2] Fisher, A. and Margolis, J. Unlocking the Clubhouse: The Carnegie Mellon Experience. SIGCSE Bulletin Special Issue on Women and Computing, Vol. 34(2), June 2002.
- [3] Beyer, Sylvia and Rynes, Kristina and Perrault, Julie and Hay, Kelly and Haller, Susan, "Gender differences in computer science students", *Proc. SIGCSE 03*, Reno, NV, 2003, 49-53.
- [4] Cohoon J.M., Gendered Experiences of Computing Graduate Programs. *Proc. SIGCSE 07*, Covington, KY, March 2007, 546-550.
- [5] Klawe, A. Girls, Boys and Computers. SIGCSE Bulletin Special Issue on Women and Computing. 34(2), June 2002, 16-17
- [6] Kumar, A.N, "The Effect of Using Problem-Solving Software Tutors on the Self-Confidence of Female Students", *Proc. SIGCSE 08*, Portland, OR, Mar 2008, 523-527.
- [7] Madigan, E.M., Goodfellow, M., Stone, J. A., "Gender, Perceptions, and Reality: Technological Literacy Among First-Year Students", *Proc. SIGCSE 07*, Covington, KY, March 2007, 410-414.
- [8] Imhof, M, Vollmeyer, R. and Beierlein, C., "Computer use and the gender gap: The issue of access, use, motivation, and performance", *Computers in Human Behavior*, 23 (6), Nov. 2007, 2823-2837.
- [9] Kao, G., "Asian Americans as model minorities? A look at their academic performance", *American Journal of Education*, 103, 1995, 121-159.
- [10] Varma, R., "Making Computer Science Minority-Friendly", *Communications of the ACM*, 49 (2), Feb 2006, 129-134.
- [11] Kumar, A.N, "Do female students feel differently than male students about using software tutors?", *Proc. FIE 06*, San Diego, CA, Oct 2006, Session S3G.
- [12] Kumar, A.N, "Female Students Assess Software Tutors More Positively Than Male Students", *Proc. FIE 08*, Saratoga Springs, NY, Oct 2008, Session S4F.
- [13] Kumar, A.N, "Patterns in Student Assessment of Problem-Solving Software", *Proc. FIE 09*, San Antonio, TX, Oct 2009, Session M2F.
- [14] Sears, J. (1999). Teaching queerly: Some elementary propositions. In W. J. Letts & J. Sears (Eds.), *Queering elementary education: Advancing the dialogue about sexualities and Schooling* (97-110). Lanham, MD: Rowman & Littlefield, Inc.