6. Results

As shown in Table 5.1, the question #5 in Test-A and Test-B were not exactly identical. A post-evaluation analysis indeed reviled that the students who took Test-B made more errors than those who took Test-A on the question #5, hence where was a main effect in the test version in the pre-test: *t*(50)=2.32; p=0.03. When we excluded the question #5 from the analysis (both in pre- and post-tests) the main effect in the test version disappeared. Hence, in the following analyses were done on the pre- and post-tests excluding question #5.

6.1. Strategy Used

Even assigned to a backward tutoring condition, some participants used forward chaining strategy for some test items, and *vice versa*. Figure 6.1 shows a number of proofs written in a discrepant strategy to the assigned tutoring condition (e.g., a participant in BC condition used forward chaining).

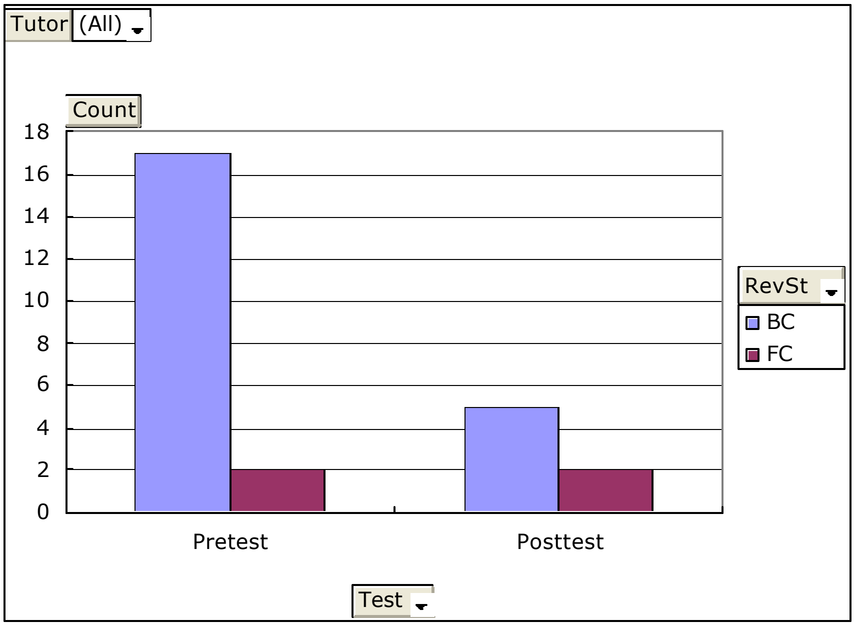


Figure 6.1: Number of Proof written in Opposite Strategy (Max=52)

As shown in the figure, 17 out of 52 (33%) of the proofs in BC tutor condition were written in forward chaining. There were ten (out of 26) participants who committed those incompatible proofs in BC. Three of those ten participants plus one new participant in BC condition used a discrepant strategy on the post-test. There was only one participant in FC tutor condition who kept using a discrepant strategy on both pre- and post-tests.

These results suggest that the participants were familiar with forward chaining prior to the study. An unofficial interview for some participants identified that they were taught forward chaining when they learned geometry theorem proving in a high school.

6.2. Learning Time

During the tutoring sessions, both tutor conditions spent almost the same amount of time for each of the problems. Figure 6.2 shows the average time spent on each problem comparing BC and FC tutor groups. A double asterisk (**) shows that the difference is statistically significant (*p*<0.05), whereas a single asterisk (*) shows a moderate difference (*p*<0.1).

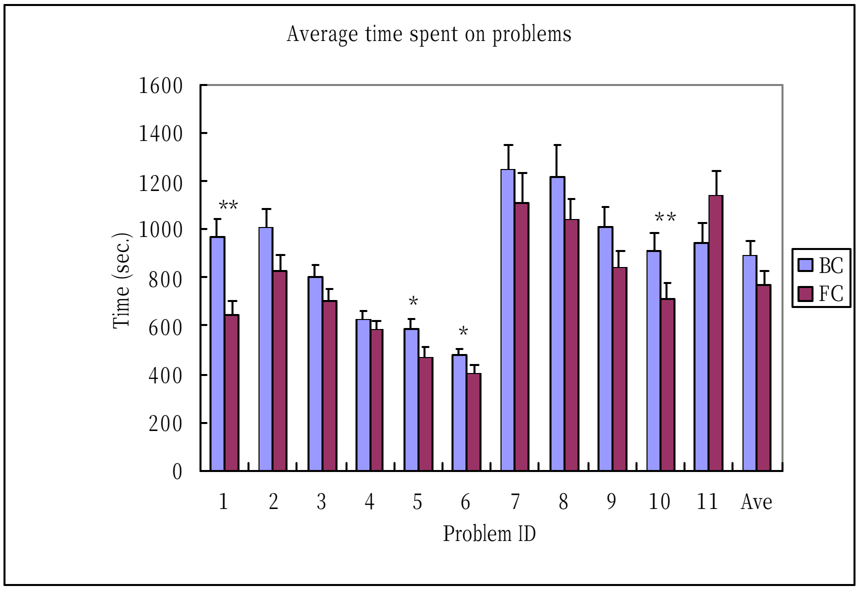


Figure 6.2: Average Time spent on a Problem

For the first two problems, the tutor, both BC and FC, provided fully proactive scaffolding that showed every single inference steps, and the participants merely watched the tutor's performance and clicked [OK] button to proceed the steps. Since backward chaining requires more steps to be performed than forward chaining, it took considerably longer for BC condition to go through first two problems.

6.3. Pre- and Post-Test Scores

Even though Test-A and Test-B are designed to be isomorphic, there was a slight difference in the length of proofs between those two tests. Also, BC tended to be longer than FC, because only BC must write givens as a part of the proof (those are written in the first few rows in the test for the FC condition). Hence, we used a ratio of correct responses as a score of the test.

For the fill-in-blank test items, a number of correct answers were counted. For the proof-writing items, a number of correct proof statements (both on-path and off-path) were counted. The score of the test was then calculated as a ratio of correct responses to the full scores.

Figure 6.3 shows the test scores on both pre- and post-tests across the tutor and the test-version conditions. Notice that the proof-writing item #5 was excluded from the analysis.

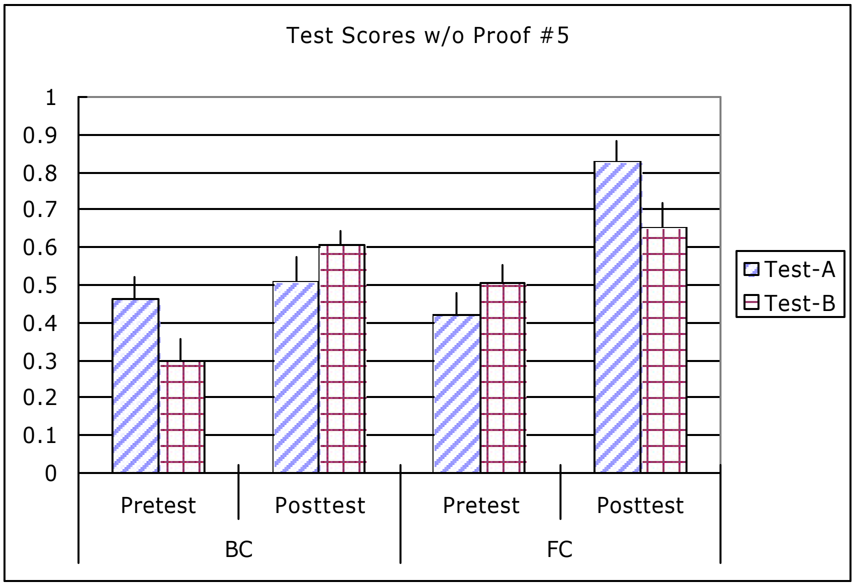


Figure 6.3: Pre- and Post-test Scores

There was no main effect either in the tutor or the test version in Pre-test scores. The interaction between the tutor and the test version was significant: *F*(1, 48)=5.18; *p*=0.03. On the other hand, in the post-test, there was a main effect on the tutor: *F*(1,48)=10.13; *p*<0.01. The difference in the test version was not significant.

A regression analysis revealed that the multiple regression equation of the post-test score upon the pre-test score and the tutor condition was:

$$\text{Post-test} = 0.52 * \text{Pre-test} - 0.14 \text{ (if BC)} + 0.50$$

The adjusted post-test scores were 0.58 and 0.72 for BC and FC conditions. The difference is equivalent to the effect size of 0.72.

6.4. Learning on the Postulates

This section summarizes participants' learning on the geometric postulates. There were 11 postulates used in the study. We first show how participants improve their performance on postulate applications. We then show how they changed their skills on postulate applications between pre- and post-tests.

6.4.1. Learning Curve

How could the participants improve their skills in applying postulate? AGT measured all sort of participants activities made during the tutoring sessions. Here we show the time and the accuracy of postulate applications. Figure 6.4 shows an average time, in millisecond, to apply a postulate. Figure 6.5 shows an average number of errors made with in a single postulate application. Both graphs show an aggregated average across all participants and all postulates in both tutor conditions. The X-axis of both graphs is the occurrence of postulate applications. Hence the left most data point in Figure 6.4, for example, show an average time taken to apply a postulate at the first time measured across all participants and all postulates.

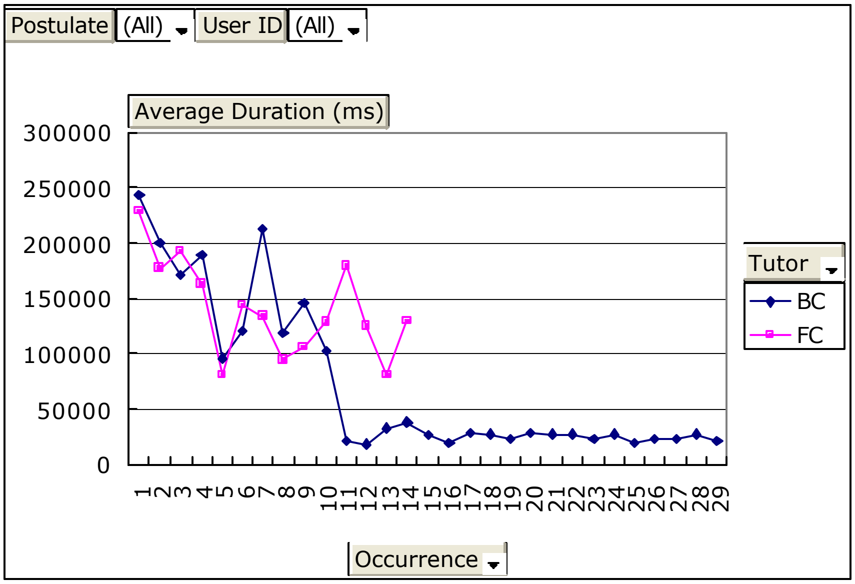


Figure 6.4: Average Duration (millisecond) for Postulate Applications [Drop GIVEN]

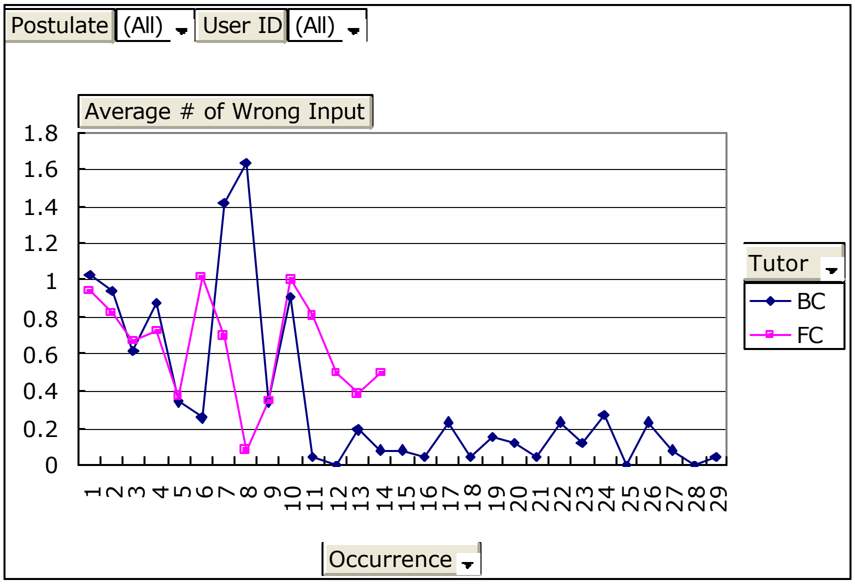


Figure 6.5: Average number of errors made during single postulate application

As shown in the figures, both tutor conditions showed a similar pattern in learning postulate applications.

6.4.2. Pre- and Post-test difference

Both BC and FC showed the same pattern in postulate learning. First, there was no difference in the accuracy of postulate applications between FC and BC conditions in both pre- and post-tests. Figure 6.6 shows the comparison between BC and FC on pre- and post-test in the accuracy of application of each postulate (ratio of correct applications to the occurrence of applications). A single asterisk shows a statistically marginal difference. All other differences were not statistically significant.

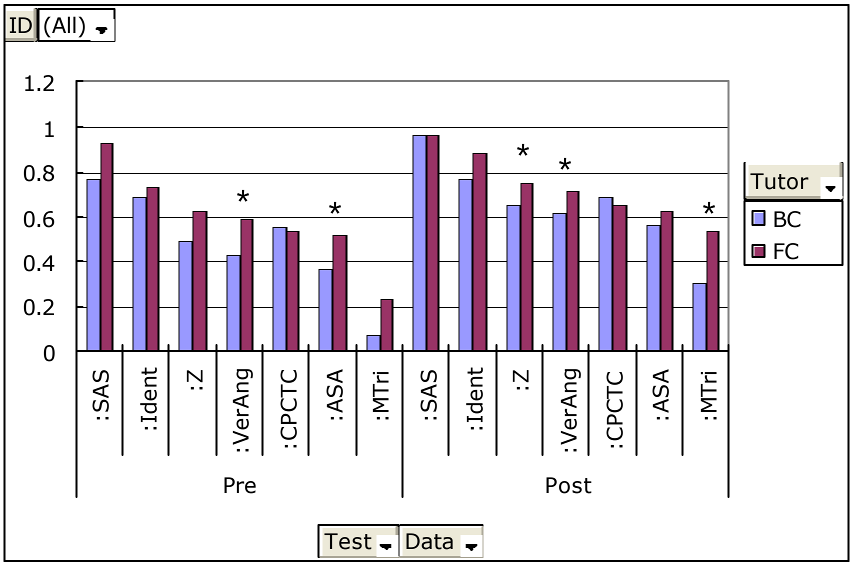


Figure 6.6: Difference between Tutor Conditions in Accuracy of Postulate Applications

Second, both BC and FC improved accuracy of postulate applications from the pre-test to the post-test. Figure 6.7 shows the comparison of the accuracy in postulate applications between pre- and post-tests for each tutor conditions. A double asterisk shows that the difference was statistically significant and a single asterisk shows a marginal difference. As can be seen in the figure, both tutor conditions showed significant improvements for most of the postulates.

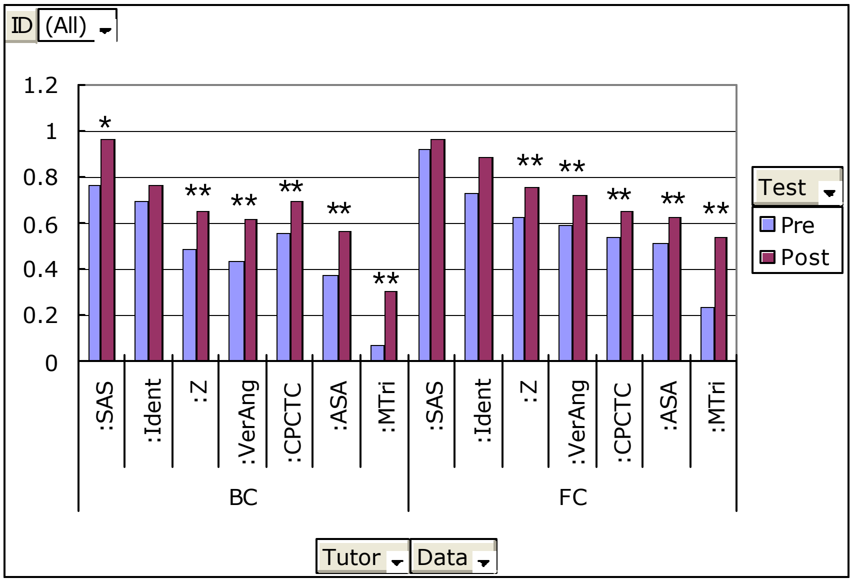


Figure 6.7: Difference in Accuracy of Postulate Applications between Tests

6.5. Performance on Proof Writing

The overall post-test score analysis showed that the BC condition made more errors on proof writing. This section provides a detailed analysis on the students' performance on proof writing.

6.5.1. Type of Erroneous Proofs

On the post-test, BC subjects in the BC condition wrote more incorrect proofs than FC condition. Figure 6.8 shows the ratio of correct and incorrect proofs written by the subjects in each tutor condition.

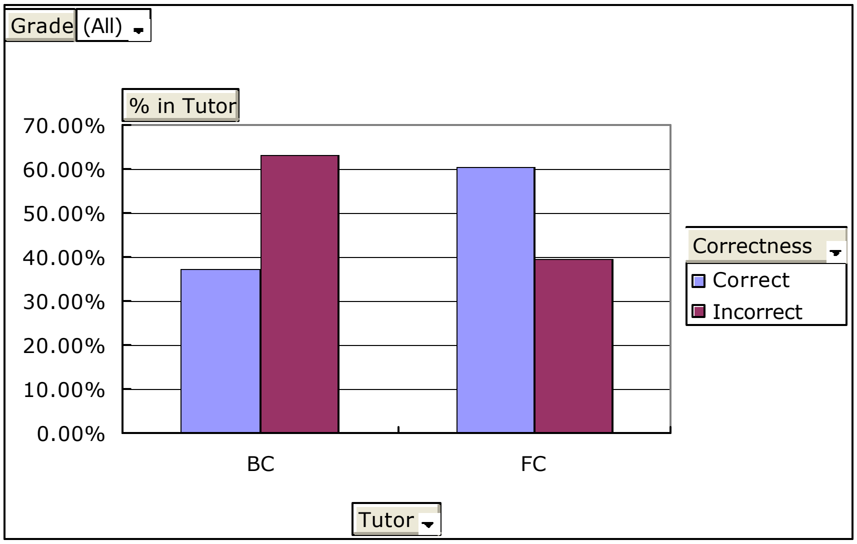


Figure 6.8: Ratio of correct and incorrect proofs in each tutor condition

What types of incorrect proofs do the participants wrote? To answer for this question, we compared the proofs written by participants regarding the strategy they applied. Figure 6.9 shows the ratio of each type of the incorrect proofs across the strategy taken. As seen in the figure, when participants applied backward chaining, they were more like to get stuck and left the proof intact without any attempt.

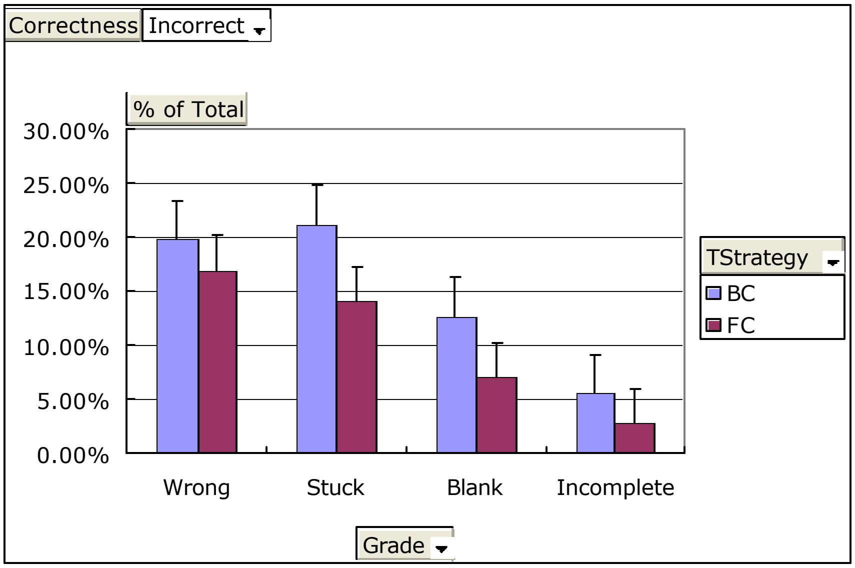


Figure 6.9: Types of incorrect proofs

The next question is then why backward chaining induced more erroneous proofs. The next sections provide an account for this inquiry.

6.5.2. Commitment to Errors

To see how the participants committed errors in incorrect proofs, we have coded all of the proof statements appeared in all three proof-writing items in the post-test (N=717). Among those 717 proof statements, 79 (11%) were coded as *incorrect* proof statements, which by definition include a false proposition, a false justification, and/or false premises. Remaining 638 (89%) proof statements were coded as *correct* both on-path and off-path proof steps. This is another support that the participants did actually grasp the concept on correct postulate application.

In those 79 incorrect proof statements, we have compared how the usage of proposition, justification and premises differ in backward and forward chaining. More precisely, we compared the occurrence of a correct, inappropriate, and wrong usage on proposition, justification, and premises. Inappropriate propositions are those that are true in the given problem configuration, but not a part of the correct proof. False proposition are those that do not hold in the problem configuration. Similarly, inappropriate justifications are those that could be justify the target proposition (i.e., its consequence unifies with the proposition to be justified), but do not indeed appear in the correct proof. Inappropriate premises are those that could be another way to support a justification, but it could not indeed a part of proof (e.g., different combination of side-angle-side for triangle congruence). The premises that do not support the justification are coded as wrong. The correctness of proposition, justification, and premises can be combined arbitrary. For example, a proof statement is coded as incorrect when its proposition and justification are both correct, but has false premises.

Figure 6.10 shows the statues on the usage of propositions. The graph shows a ratio of each type of commitment to the total number of occurrence. A correct use of proposition was coded either “Goal” or “On-path.” “Goal” referred to the proof statements that attempt to justify the top-level goal to prove, whereas “On-path” referred to the proof statements that have a propositions hat are a part of the correct proof. An inappropriate use of proposition was coded as “Off-path.” The false use of proposition was coded as “Wrong.”

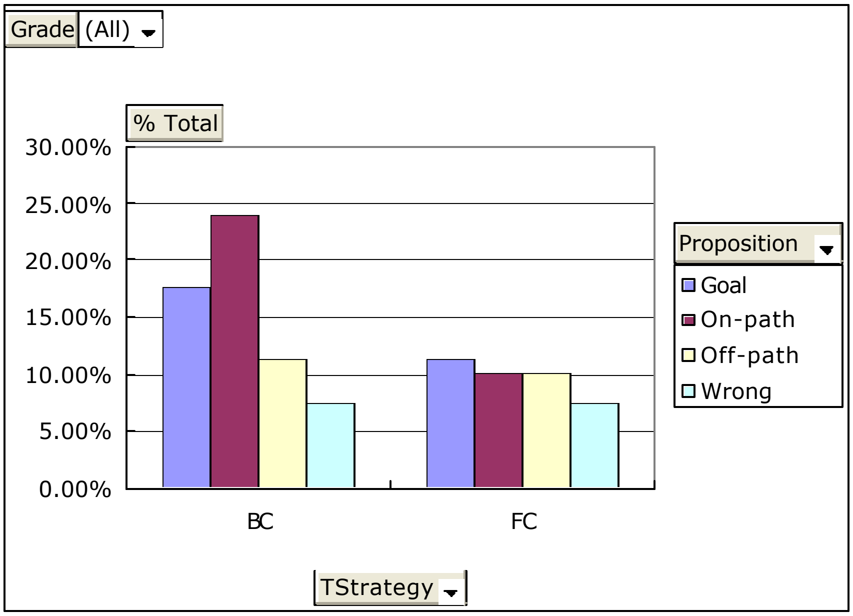


Figure 6.10: Usage of propositions

To test the difference in the usage of propositions, we built a 2 x 4 contingency table (Table 6.1); two rows of Forward and Backward chaining, and four columns of Goal, On-path, Off-path, and Wrong usage of the proposition. A Chi-Square test on the 2 x 4 contingency table showed no significant difference in the use of propositions.

Table 6.1: A 2 x 4 Contingency Table on the usage of Propositions

TStrategy * Proposition Crosstabulation

			Proposition				Total
			0	1	2	Goal	
TStrategy	BC	Count	6	9	19	14	48
		Expected Count	7.3	10.3	16.4	14.0	48.0
	FC	Count	6	8	8	9	31
		Expected Count	4.7	6.7	10.6	9.0	31.0
Total		Count	12	17	27	23	79
		Expected Count	12.0	17.0	27.0	23.0	79.0

Figure 6.11 shows the status on the use of justifications. Similar to the analysis on the propositions, justifications were coded as "On-path" when they were used in the same way as the ones in the correct proof. Sound justifications that were not in the correct proof were coded as

"Off-path." Invalid justifications were coded as "Wrong." When the justification were not provided at all, they were coded as "Blank."

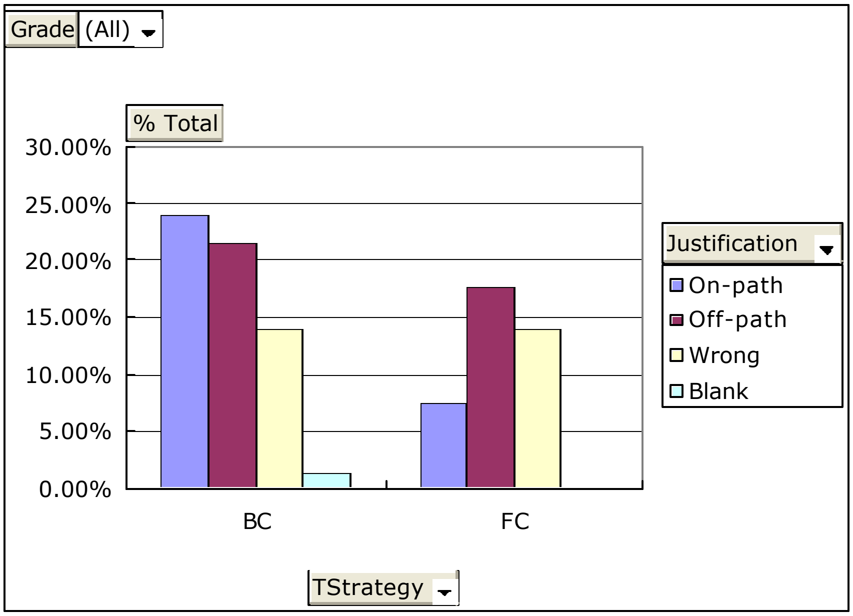


Figure 6.11: Usage of Justificaiton

A Fisher's Exact Test a 2 x 4 contingency table (Table 6.2) did not show a significant difference in the usage of justification between forward and backward chaining.

Table 6.2: A 2 x 4 Contingency Table on the use of Justification

TStrategy * Justification Crosstabulation

			Justification				Total
			-1	0	1	2	
TStrategy	BC	Count	1	11	17	19	48
		Expected Count	.6	13.4	18.8	15.2	48.0
	FC	Count	0	11	14	6	31
		Expected Count	.4	8.6	12.2	9.8	31.0
Total		Count	1	22	31	25	79
		Expected Count	1.0	22.0	31.0	25.0	79.0

So far, there was no significant difference in the use of proposition and justification. There was, however, a significant difference in the use of premises between forward and backward chaining. Figure 6.12 shows the status on the use of premises. As shown in the figure, backward chaining did make more commitment on the “Blank” premises, namely, backward chaining were more likely to fail to provide correct premises.

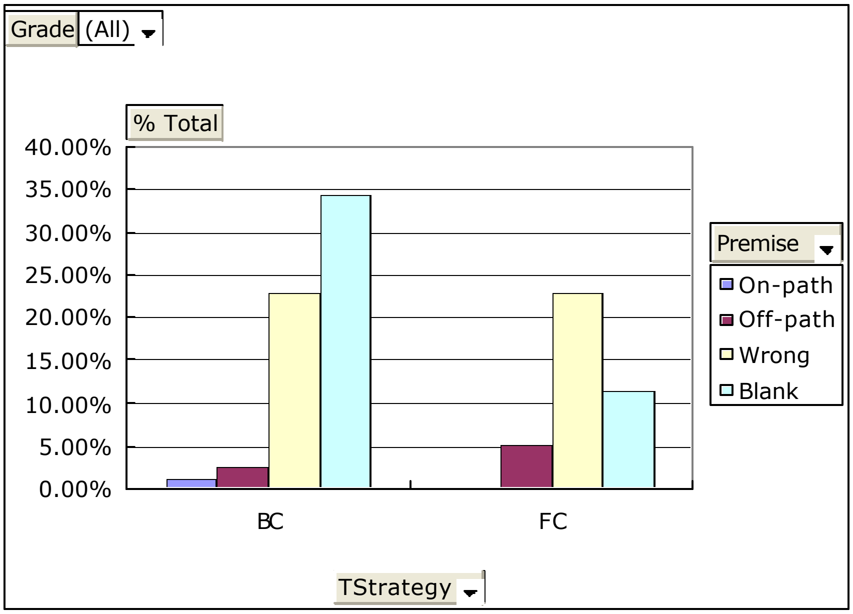
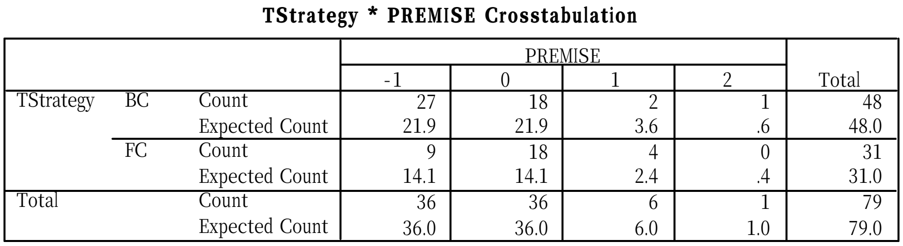


Figure 6.12: Usage of Premise

A Fisher’s Exact test on a 2 x 4 contingency table (Table 6.3) actually revealed a significant difference: Fisher’s Exact Test = 7.25; *p*=0.04.

Table 6.3: A 2 x 4 Contingency Table on the use of Premises



TStrategy * PREMISE Crosstabulation

			PREMISE				Total
			-1	0	1	2	
TStrategy	BC	Count	27	18	2	1	48
		Expected Count	21.9	21.9	3.6	.6	48.0
	FC	Count	9	18	4	0	31
		Expected Count	14.1	14.1	2.4	.4	31.0
Total		Count	36	36	6	1	79
		Expected Count	36.0	36.0	6.0	1.0	79.0

To see the use of premises in details, we have conducted a Chi-Square test on a 2 x 2 contingency table whose column consists of one category of premise use and aggregation of remaining categories. The analysis revealed that BC tended to leave the premises blank more often than FC (Chi-Square = 5.63, *p*=0.02). BC also tended to use wrong premises more often than FC (Fisher’s exact test = 3.21, *p*=0.06). There was no significant difference in the use of neither inappropriate nor correct premises between FC and BC.

7. Discussion

After proving 11 geometry theorems under an intensive aid from AGT, the participants in the forward chaining and the backward chaining tutor conditions did show different performance on proof writing in the post-test.

7.1. Impact on the Difference in the Proof Strategy

That the performance of participants in the forward chaining tutor condition outperformed the backward chaining condition is consistent with what Anderson *et al*. observed in their geometry study (Anderson et al., 1993). Backward chaining certainly is a challenging skill for students to learn.

The participants apparently had became familiar with forward chaining, which can be seen in that fact that ??% of the participants in the backward chaining condition applied forward chaining. This may be an account for the high scores in forward chaining condition. << so what? >>

Both condition had the similar pattern in proof writing.

7.2. Difficulty of Thinking Backwards

We have also identified a potential source of the difficulty for backward chaining. The most striking finding is that the participants in the backward chaining condition tend to get stuck on proving premises even for their sound postulate applications. In other words, it seems to be difficult for students to specify “subgoals” to justify a current goal in mind. That the participants in both backward chaining and forward chaining conditions equally