AGT Evaluation (2): Comparison with the Test Scores

Noboru Matsuda
University of Pittsburgh
June 24, 2004

1. Test Scores

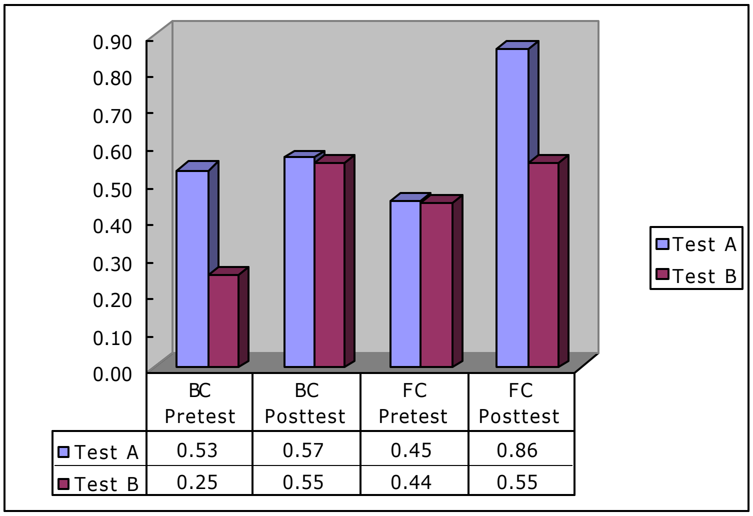


Figure 1: Test Scores

- Overall, FC outperformed BC in the posttest, but they tied in the pretest.
- Those who took Test-B as a pretest showed progress on the post-test, but not those who took Test-A first.
- In BC, there was a difference between Test-A and B scores in the pretest, but no such difference in posttest.
- In FC, there was a difference between Test-A and B scores in the posttest, but no such difference in pretest.
- Among those who took Test-B as a pretest, the FC group showed more learning gain on the pre-test (i.e., Test-A) than the BC group.
- In BC, Test-A does not include much of what AGT addressed during tutoring, because the scores of Test-A is equal regardless of its usage (i.e., Test-A-pretest tied Test-A-posttest).

- In FC, Test-B does not include much of what AGT addressed during tutoring, because the score of Test-B is equal regardless of its usage.
- In BC, AGT addressed something that is closely related to Test-B. They correspond to proof writing problems, not fill-in-a-blank.
- In FC, AGT addressed something that is closely related to Test-A. They correspond to proof writing problems, not fill-in-a-blank.

2. Overall comparison

I have yet to know how to deal with the difference in BC pretest.

3. Interaction between Learning Gain and Test Items

Those who took Test-B as a pretest showed progress on the post-test, but not those who took Test-A first.

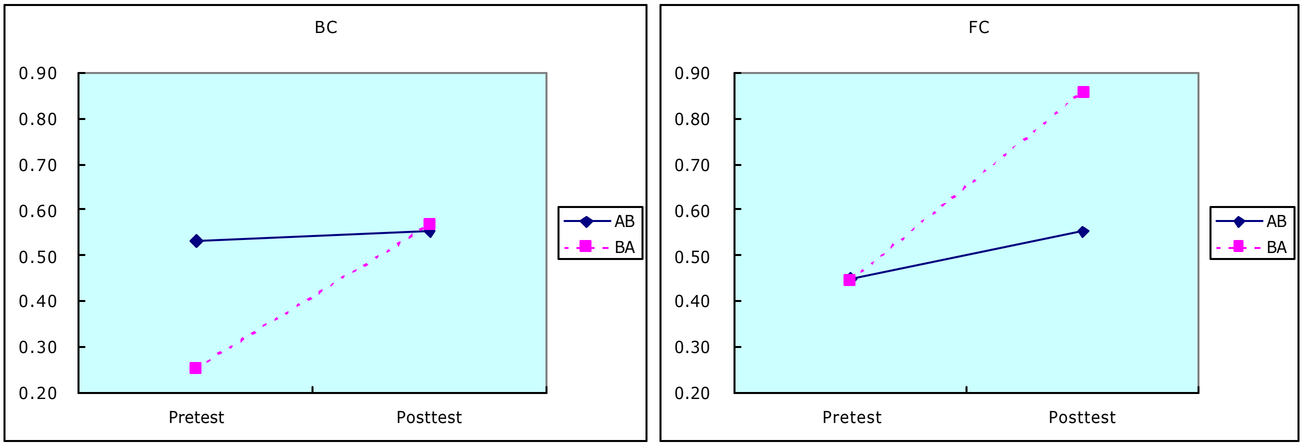


Figure 2: Learning Gain in BC and FC conditions

As shown in Figure 2, there is an interaction between the order of test items and learning gain in both tutor conditions. Namely, those who took Test-B as a pretest (the BA group) shows bigger learning gain than the ones who took Test-A as a pretest regardless of the type of tutor. The interactions are significant.

BC: Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.877[a]	3	.292	7.547	.000
Intercept	11.779	1	11.779	304.219	.000
ORDER	.233	1	.233	6.015	.018
TEST	.364	1	.364	9.389	.004
ORDER * TEST	.280	1	.280	7.238	.010
Error	1.858	48	.039		
Total	14.514	52			
Corrected Total	2.735	51			

[a] R Squared = .321 (Adjusted R Squared = .278)

FC: Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.487[a]	3	.496	9.840	.000
Intercept	17.234	1	17.234	342.046	.000
ORDER	.297	1	.297	5.892	.019
TEST	.873	1	.873	17.326	.000
ORDER * TEST	.318	1	.318	6.303	.015
Error	2.418	48	.050		
Total	21.140	52			
Corrected Total	3.906	51			

[a] R Squared = .381 (Adjusted R Squared = .342)

The paired T-test in each group showed significant difference in Pre- (SCORE1) and Post-test (SCORE2) only those who took Test-B as a pre-test (ITEM1) in both the BC and FC tutor conditions.

BC: Paired Samples Test

			Paired Differences							
						95% Confidence Interval of the Difference				
ITEM1			Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
A	Pair 1	SCORE1 - SCORE2	-.0204	.22041	.06113	-.1536	.1128	-.334	12	.744
B	Pair 1	SCORE1 - SCORE2	-.3140	.16393	.04547	-.4131	-.2150	-6.907	12	.000

FC: Paired Samples Test

			Paired Differences							
						95% Confidence Interval of the Difference				
ITEM1			Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
A	Pair 1	SCORE1 - SCORE2	-.1028	.27187	.07540	-.2671	.0615	-1.364	12	.198
B	Pair 1	SCORE1 - SCORE2	-.4154	.20290	.05627	-.5381	-.2928	-7.382	12	.000

4. Within a Tutor between Test-Items Comparison

In BC, there was a difference between Test-A and B scores in the pretest, but no such difference in posttest. In FC, the effect went the other way around.

Independent Samples Test[a]

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
PRETEST	Equal variances assumed	1.195	.285	3.871	24	.001	.280667	.0725067	.1310202	.4303130
	Equal variances not assumed			3.871	21.481	.001	.280667	.0725067	.1300861	.4312471
POSTTEST	Equal variances assumed	.140	.712	-.159	24	.875	-.012987	.0815838	-.1813681	.1553934
	Equal variances not assumed			-.159	23.990	.875	-.012987	.0815838	-.1813718	.1553971

[a] TUTOR = BC

Independent Samples Test[a]

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
PRETEST	Equal variances assumed	.952	.339	.055	24	.956	.005183	.0938034	-.1884179	.1987834
	Equal variances not assumed			.055	22.756	.956	.005183	.0938034	-.1889792	.1993448
POSTTEST	Equal variances assumed	1.389	.250	-3.755	24	.001	-.307420	.0818787	-.4764089	-.1384302
	Equal variances not assumed			-3.755	20.241	.001	-.307420	.0818787	-.4780852	-.1367539

[a] TUTOR = FC

5. Within a Test-Item between Tutors Comparison

Among those who took Test-B as a pretest, the FC group showed more learning gain on the pre-test (i.e., Test-A) than the BC group.

ANACOVA between the tutors on the learning gain (Posttest – Pretest) with the Pretest score as a covariate shows that there is a significant difference in the learning gain between BC and FC only among those who took Test-B as a pre-test (p = .014).

Tests of Between-Subjects Effects[b]

Dependent Variable: GAIN

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.508[a]	2	.254	5.809	.009
Intercept	.563	1	.563	12.864	.002
PRETEST	.464	1	.464	10.608	.003
TUTOR	.007	1	.007	.167	.687
Error	1.006	23	.044		
Total	1.613	26			
Corrected Total	1.514	25			

[a] R Squared = .336 (Adjusted R Squared = .278)

[b] ITEM1 = A

Tests of Between-Subjects Effects[b]

Dependent Variable: GAIN

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.260[a]	2	.130	4.806	.018
Intercept	1.510	1	1.510	55.738	.000
PRETEST	.194	1	.194	7.145	.014
TUTOR	.192	1	.192	7.083	.014
Error	.623	23	.027		
Total	4.342	26			
Corrected Total	.883	25			

[a] R Squared = .295 (Adjusted R Squared = .233)

[b] ITEM1 = B

6. Within a Tutor and a Test-Item Comparison

In BC, the scores of Test-A are equal in pre- and post-test, whereas there is a significant difference in Test-B scores. In FC, the reverse occurred.

BC: Independent Samples Test

			Levene's Test for Equality of Variances		t-test for Equality of Means						
										95% Confidence Interval of the Difference	
ITEM			F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
A	SCORE	Equal variances assumed	.058	.812	-.401	24	.692	-.0334	.08321	-.20513	.13836
		Equal variances not assumed			-.401	23.991	.692	-.0334	.08321	-.20513	.13837
B	SCORE	Equal variances assumed	2.019	.168	-4.262	24	.000	-.3011	.07063	-.44684	-.15529
		Equal variances not assumed			-4.262	21.932	.000	-.3011	.07063	-.44757	-.15456

Independent Samples Test

			Levene's Test for Equality of Variances		t-test for Equality of Means						
										95% Confidence Interval of the Difference	
ITEM			F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
A	SCORE	Equal variances assumed	4.815	.038	-4.790	24	.000	-.4103	.08565	-.58702	-.23349
		Equal variances not assumed			-4.790	19.507	.000	-.4103	.08565	-.58920	-.23131
B	SCORE	Equal variances assumed	.031	.861	-1.195	24	.244	-.1080	.09038	-.29454	.07851
		Equal variances not assumed			-1.195	23.291	.244	-.1080	.09038	-.29485	.07881

So, the BC tutor did something good for Test-B, and the FC tutor did something good for Test-A. What are they?

7. Analysis on Test Items: Competence on Postulates

The scores in fill-in-a-blank test items are all same regardless of the test item and the tutor.

A pre- and post-test consists of fill-in-blank items and write-a-proof items. There are 6 blanks to fill in as a part of three different proofs. Students provided a postulate name for each of the blanks. As shown in Figure 3, in the pre-test in FC condition, there is a moderate (p=.089) difference in the number of correct fill-in-blank items between Test-A and Test-B. In the post-test in BC condition, there is a moderate (p=.073) difference in the number of correct fill-in-blank items between Test-A and Test-B. Therefore, the difference in learning gain should appear as a difference in scores of write-a-proof items.

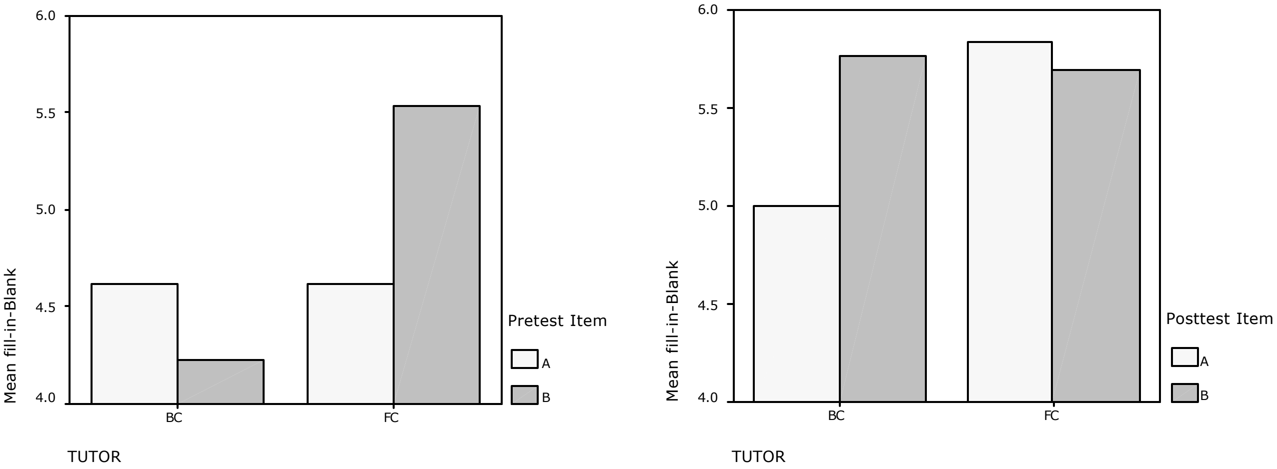


Figure 3: Number of correct fill-in-blank items (Max 6)

There are three proof problems both in a pre- and a post-test. As shown in Figure 4, in the Pre-Test in BC condition, there is a significant difference in number of correct proofs between Test-A and Test-B (p=.000). In the Post-Test in FC condition, there is a significant difference in number of correct proofs between Test-A and Test-B (p=.002).

- Those who took Test-B as a pre-test in BC condition somehow started from low score. Test-B in FC condition is as good as other tests. Proving proof problems in Test-B backwards is considerably more difficult than proving them forwards.
- FC tutor affected quite positively to prove problems in Test-A, but not in Test-B.

Figure 5 shows the comparison with proof-writing items to see the difference in proof writing for the same test items before and after the tutoring sessions:

- In BC condition, those who took Test-A as a post-test ended up with the "bottom-line" (i.e., Pre- and Post-test scores of Test-A tied). This means that they did not learn anything at all.
- Those who took Test-A as a pre-test in BC condition ended up with the same post-test score. However, given that the Test-B's pre-test score is quite low, they must have learned something.
- In FC condition, those who took Test-B as a pre-test showed significant gain on the post-test: FC tutor did teach something good to prove problems in Test-A.
- In FC condition, those who took Test-A first didn't learn at all.

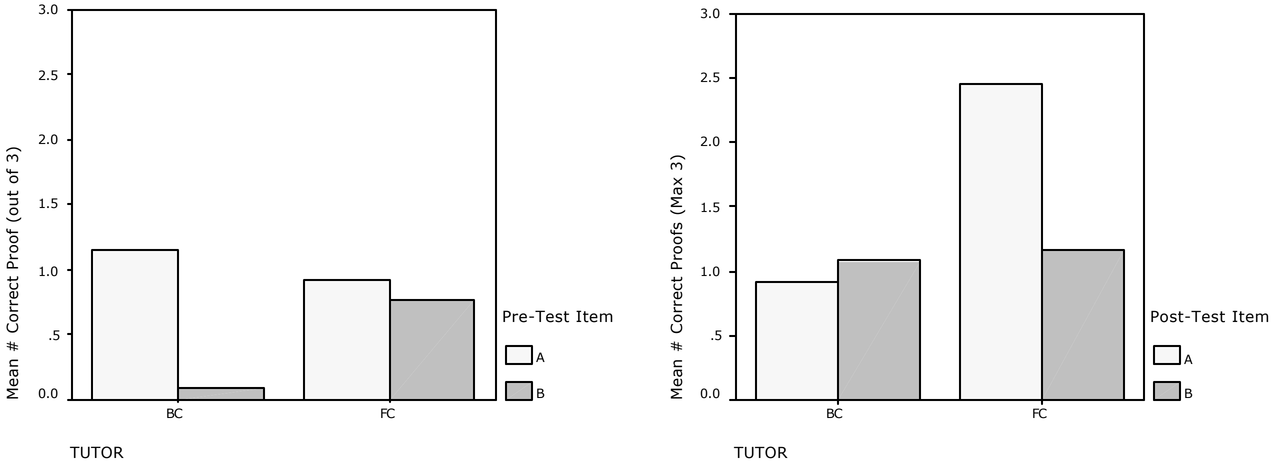


Figure 4: Number of correct proofs (Between test items comparison)

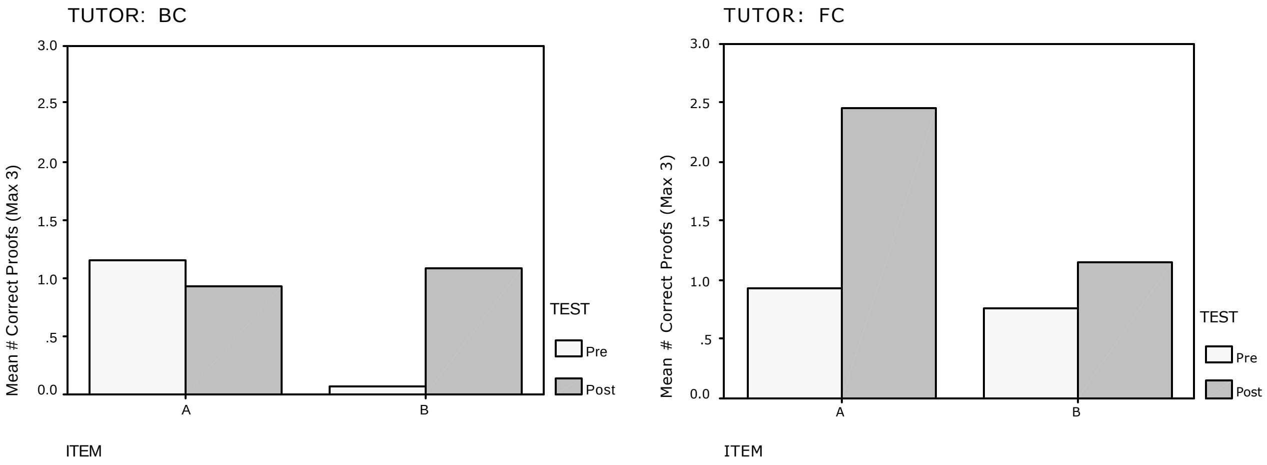
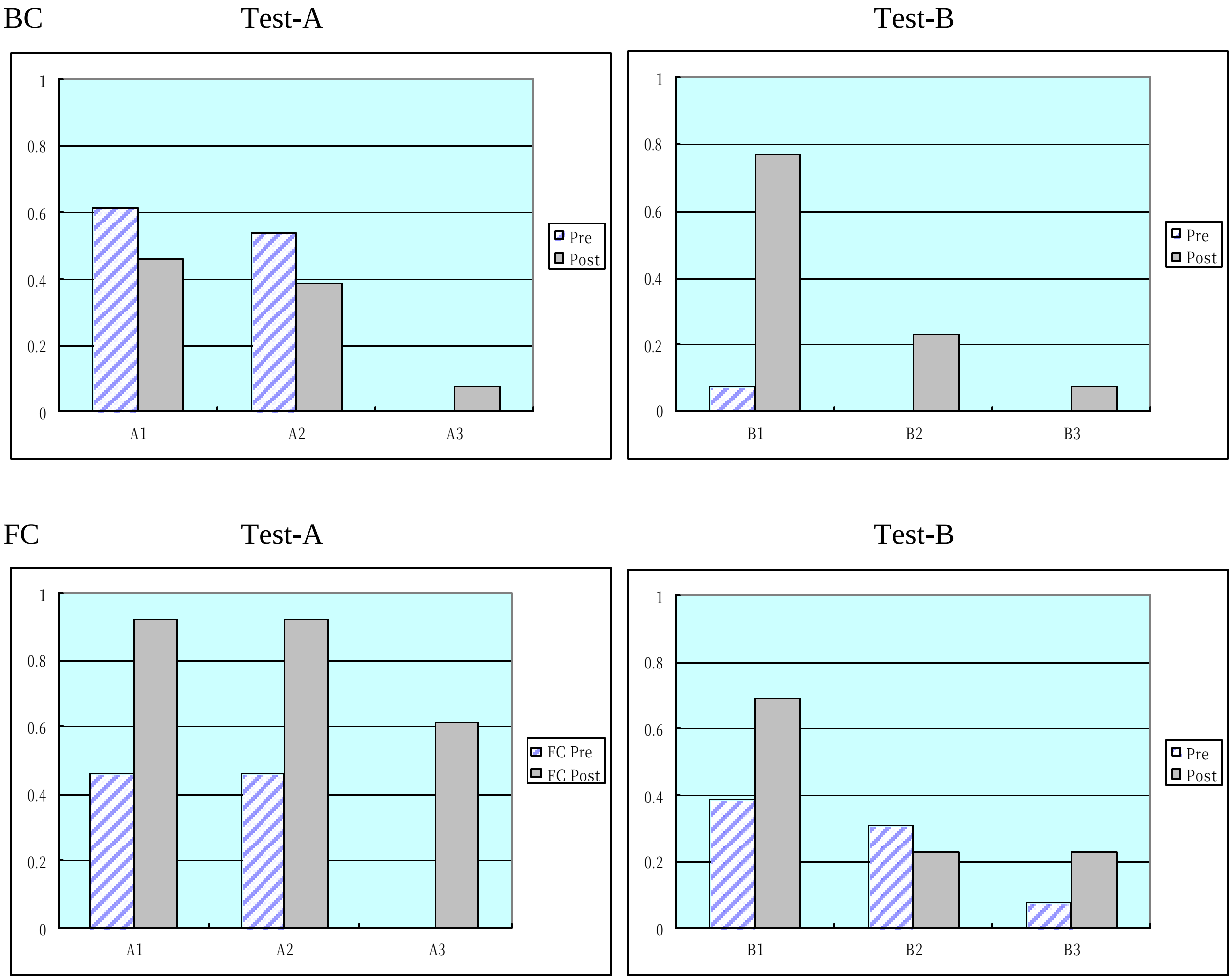


Figure 5: Number of correct proofs (Between Pre- and Post- comparison)

Figure 6 shows the average number of correct proof for each write-a-proof test item (N=13 for each). Each category shows a difference in the average number between pre- and post-test.



BC Test-A Test-B

FC Test-A Test-B

Figure 6: Comparison with improvement in average number of correct proofs