Exam Fairness: The Effect of Blue Paper Versus White Paper on Student Performance

James E. De Franceschi
Southeastern Louisiana University

Pierre L. Titard
Southeastern Louisiana University

Harold E. Davis
Southeastern Louisiana University

There is literature suggesting that white paper is too harsh and a more soothing color might be better for exams. Prior studies using various colors of paper for exams have failed to show any significant benefit of using any one color over another. This paper examines the results of five exams administered to four sections of managerial accounting students. Half the students received exams printed on white paper and the other half received exams printed on blue paper. Contrary to conventional wisdom, preliminary t-tests indicate that students receiving the white exam scored significantly higher than those receiving the blue exam.

BACKGROUND

Accounting instructors often use exams of different colors for security purposes. One of the concerns with doing this is whether the color of the exam may have an influence on the test takers performance. Shacklett, et. al. (2005) provide an extensive review of the literature on color and it suggests that color does have an effect on people. They also surmised that students taking exams on white paper would not do as well due to the glare and fatigue. Their paper then reports on the effects of using colored exams in a marketing class. They reported on 286 exam papers in five colors, including blue and white. Using t-test, they found no significant difference in test scores for any of the colors. However, blue exams had the highest average and white exams the second highest.

Titard, et. al. (2006), conducted a similar study using management accounting students. Two instructors administered over 1,200 exams using white and five other colors of paper, including blue. Scores for the white exams were compared to the scores of the colored exams using t-tests. No significant differences were found between the scores. Instructor 1 combined averages in one semester were higher on the white exams, but in two other semesters the color averages were higher. Instructor 2 reported higher combined averages for colored exams.

Comparisons between the white and specific colors or between specific colors were not performed. A study by Pinar and Fogliasso (2005) also compared scores of different color exams. They were given in three different undergraduate business classes and an MBA class. Blue was used as a control color.

Individual t-tests were run for each exam with no significant difference detected for the blue or any other color. Class level comparisons were then done. For three of the four classes and all classes combined there were no significant differences. In the fourth class, there were some significant color differences. Of interest to this study was that the white exam scores were significantly higher than the blue. Also, while not significant, the means for the white exams were higher on six of the eight exams where the colors were white and blue.

METHODOLOGY

The studies cited above showed no significant results, but mixed anecdotal evidence on blue or white exams. Instructors at a regional university in the southeast United States use alternate exams in the managerial and an accounting principles class. They do this for security purposes in classrooms where students sit very close to each other. Exams A and B are virtually the same except for color. Questions are the same on both versions of the exam, but the answers or the questions themselves may be in a different order.

In this study the exams are either white or blue. They are distributed so that alternate rows will receive one color of the exam and the other rows the other color. Students that receive a white (blue) exam for the first, third , and fifth exams receive a blue (white) one for the second and fourth exams. Students receive both colors over the course of the semester and do not feel at a disadvantage. The different colors reinforce the fact that the exams are different and prevent the students' swapping an exam across the row. It also keeps the instructor from distributing the wrong exam to a row. By alternating exam color each test, we felt that this would negate the effect of good (poor) students always getting the same color.

RESULTS

Two instructors administered exams using white and blue as the alternate colors. The data from five semesters were used for this study. The scores from 1,405 exams (696 blue, 709 white) were entered and a t-test was run. Results are shown in Table 1, column 2. The white scores were significantly better than the blue. Since instructor two only gave one exam that was blue and white in each section and his exams were a different format than instructor one's, we decided to delete them from the study and run another t-test (Table 1, column 3). We then deleted the accounting principles scores to just compare the managerial accounting (Acct. 225) results (Table 1, column 4). Finally, we deleted one semester of the Acct. 225 scores because the first exam used different colors. This leaves column 5 of table 1 which shows the results for one semester in which all five exams are blue and white. All the results show the scores from the white exams being significantly higher than the blue scores.

TABLE 1
INITIAL T-TEST

EXAM COLOR	INSTRUCTORS	INSTRUCTOR	INSTRUCTOR 1	INSTRUCTOR 1
	1,405 EXAMS	1,188 EXAMS	999 EXAMS (ACCT. 225 ONLY)	600 EXAMS (ACCT. 225 SP 2006)
	MEAN	MEAN	MEAN	MEAN
WHITE	71.06	71.32	71.34	72.34
BLUE	69.64	69.28	68.90	68.63
T-TEST	.09	.025	.013	.005

The remainder of the study just reports on the one semester of Accounting 225 that is shown in column 5 of Table 1. We performed t-tests on each individual exam per class and all classes combined.

The individual classes were also tested. The results are shown in Table 2. Most individual class exams show no significant difference, but those that do show a difference favor the white exams.

TABLE 2
T-TEST - ACCOUNTING 225 – 1 SEMESTER

SECTION	EXAM 1	EXAM 2	EXAM 3	EXAM 4	EXAM 5	TOTAL
5	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
6	.02 white	n.s.	n.s.	n.s.	n.s.	.015 white
7	n.s.	.053 white	n.s.	n.s.	.029 white	.055 white
8	n.s.	n.s.	n.s.	n.s.	.021 white	n.s.
TOTAL	n.s.	n.s.	n.s.	n.s.	.006 white	.005 white

Exam 1 shows a significant difference only in section 6 but not in total. Exam 2 shows a significant difference only in section 7 but not in total. Exams 3 and 4 show no significant differences. Exam 5 shows a significant difference in sections 7 and 8 and in total. Section 6 shows a significant difference only on exam 1, but also in total. Section 7 shows a significant difference on exams 2 and 5, and in total. Since students who get the white exam 2 get a blue exam 5, we felt the difference shouldn't be due to student ability.

Although we felt that alternating the color of test between students should eliminate factors other than color, we felt it best to run an ANOVA and control for other possibilities. We included the students' GENDER, grade point average (GPA), and SECTION along with COLOR in the analysis of each exam. The summarized results for color are shown in Table 3. Although the white averages are higher for every exam, when the other variables are included in an ANOVA, there is no significant difference due to paper color for any of the individual exams.

TABLE 3
SCORE * COLOR ANOVA

EXAM	MEAN				
COLOR	EXAM 1	EXAM 2	EXAM 3	EXAM 4	EXAM 5
WHITE	73.59	68.06	66.44	71.14	80.13
BLUE	71.89	65.9	64.26	69.84	72.56
SIG.	.898	.321	.119	.338	.290

Other observations: none of the included variables were significant in exams 1 or 2; GENDER was significant in exam 3 (.008); SECTION was significant in exam 4 (.042) and exam 5 (.095); and GPA was significant in exam 3 (.054), exam 4 (.034), and exam 5 (.034). We will comment on this later.

The results above are for individual exams. When we ran an ANOVA including all five exams, with EXAM as one of the variables, we received an error because the GPA we were using was not grouped and we had a lot of observations. So, we grouped the GPA into another variable you see in the results called **GPA2**. The grouping is as follows:

1 = 0.000 - 0.499
2 = 0.500 - 0.999
3 = 1.000 – 1.499
4 = 1.500 – 1.999
5 = 2.000 – 2.499
6 = 2.500 – 2.999
7 = 3.000 – 3.499
8 = 3.500 – 4.000

TABLE 4
ANOVA- ALL EXAMS

Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	46611.250[a]	15	3107.417	17.036	.000
Intercept	385283.287	1	385283.287	2112.284	.000
COLOR	1145.063	1	1145.063	6.278	.012
GENDER	397.767	2	198.883	1.090	.337
EXAM	9712.492	4	2428.123	13.312	.000
SECTION	1289.613	3	429.871	2.357	.071
GPA2	28315.808	5	5663.162	31.048	.000
Error	111447.165	611	182.401		
Total	3263065.000	627			
Corrected Total	158058.415	626			

a. R Squared = .295 (Adjusted R Squared = .278)

TABLE 5

SCORE * gpa2

SCORE

gpa2	Mean	N	Std. Deviation
3.00	53.60	5	17.344
4.00	62.51	67	17.669
5.00	64.83	187	14.696
6.00	68.55	164	14.539
7.00	75.78	106	12.653
8.00	84.36	98	10.512
Total	70.37	627	15.890

TABLE 6

SCORE * EXAM

SCORE

EXAM	Mean	N	Std. Deviation
1	72.72	139	11.719
2	67.01	129	14.000
3	65.38	125	17.708
4	70.49	113	18.321
5	76.31	121	15.253
Total	70.37	627	15.890

Table 4 shows the results when all exams are included in the ANOVA. That GPA2 is significant is expected. Better students will have higher test grades, as shown in Table 5.

The change in means for the exams, as shown in Table 6, may be due to poorer students dropping the class after exam 3. In addition, those with lower averages may study more for the last exam in an effort to raise their final grade.

Table 7 shows the means of the four sections in the study. Sections 6 and 7 were at 11:00 and 12:30, respectively. Sections 5 and 8 were at 9:30 and 3:30, respectively. The lower scores in the 3:30 section are not surprising, as these sections traditionally do poorer. It could be from student or teacher fatigue toward the end of the day. The 9:30 difference may be due to students' not being alert and ready for class.

TABLE 7

SCORE * SECTION

SCORE

SECTION	Mean	N	Std. Deviation
05	68.07	128	16.391
06	72.33	186	15.475
07	73.33	141	14.132
08	67.54	172	16.697
Total	70.37	627	15.890

Table 8 shows that the white exams had a higher mean score than the blue exams. This difference is significant at the .012 level, indicating that there is an advantage for the students who had exams administered on white paper.

TABLE 8

SCORE * COLOR

SCORE

COLOR	Mean	N	Std. Deviation
White	71.78	314	15.385
Blue	68.96	313	16.283
Total	70.37	627	15.890

We did run another ANOVA in which all of the possible interactions were included to see how the significance of the main variables would hold up. The results are shown in Table 9. If a variable such as **GPA2** was strong, its significance would not be altered by the inclusion of other interactions. We do see that in light of all possible interactions **GPA2**, **Exam**, and **Section** remain significant and **Gender** does not, which is consistent with our previous results. However, **COLOR** has become insignificant in the all-interaction model, below.

TABLE 9 – ALL INTERACTIONS MODEL

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	104836.331[a]	324	323.569	1.836	.000
Intercept	689163.470	1	689163.470	3910.545	.000
GPA2	12929.662	5	2585.932	14.673	**.000**
COLOR	255.806	1	255.806	1.452	**.229**
EXAM	1744.377	4	436.094	2.475	**.044**
GENDER	165.063	2	82.532	.468	**.627**
SECTION	1669.906	3	556.635	3.159	**.025**
GPA2 * COLOR	581.308	4	145.327	.825	.510
GPA2 * EXAM	3114.134	18	173.007	.982	.481
COLOR * EXAM	506.122	4	126.531	.718	.580
GPA2 * COLOR * EXAM	3453.244	16	215.828	1.225	.247
GPA2 * GENDER	843.582	4	210.895	1.197	.312
COLOR * GENDER	285.190	1	285.190	1.618	.204
GPA2 * COLOR * GENDER	124.611	4	31.153	.177	.950
EXAM * GENDER	1418.098	8	177.262	1.006	.432
GPA2 * EXAM * GENDER	2213.167	16	138.323	.785	.703
COLOR * EXAM * GENDER	607.493	4	151.873	.862	.487
GPA2 * COLOR * EXAM * GENDER	1699.908	15	113.327	.643	.838
GPA2 * SECTION	3474.310	12	289.526	1.643	.079
COLOR * SECTION	897.203	3	299.068	1.697	.168
GPA2 * COLOR * SECTION	1806.199	11	164.200	.932	.510
EXAM * SECTION	2081.858	12	173.488	.984	.464
GPA2 * EXAM * SECTION	5340.961	45	118.688	.673	.946
COLOR * EXAM * SECTION	2806.399	12	233.867	1.327	.202
GPA2 * COLOR * EXAM * SECTION	10199.965	35	291.428	1.654	.014
GENDER * SECTION	1279.375	4	319.844	1.815	.126
GPA2 * GENDER * SECTION	1110.487	6	185.081	1.050	.393
COLOR * GENDER * SECTION	528.761	3	176.254	1.000	.393
GPA2 * COLOR * GENDER * SECTION	374.635	3	124.878	.709	.548
EXAM * GENDER * SECTION	987.495	12	82.291	.467	.933
GPA2 * EXAM * GENDER * SECTION	1522.518	18	84.584	.480	.965
COLOR * EXAM * GENDER * SECTION	2575.931	12	214.661	1.218	.269
GPA2 * COLOR * EXAM * GENDER * SECTION	299.687	6	49.948	.283	.945
Error	53222.083	302	176.232		
Total	3263065.000	627			
Corrected Total	158058.415	626			

a. R Squared = .663 (Adjusted R Squared = .302)

CONCLUSIONS

Although the ANOVA with all exams included shows a significant difference in tests scores between those with exams given on white paper and those given on blue paper, this becomes insignificant when all possible interactions are included. Based on these findings, we would conclude that using white or blue paper to give an exam has little or no affect on the outcome of the students' performance in a class. Other variables or interactions between variables may have more significance on the outcome.

Certain caveats must be noted. First, only one course was used in the study and other subjects may be more sensitive to colors. All exams were multiple choice and from one teacher. Further study should include multiple teachers giving standardized exams so that teacher effect could be included in the study. The exams in this study were printed on white paper or Sparco brand, number 05121 blue paper. Another shade of blue paper may produce different results.

REFERENCES

Pinar, M., and Fogliasso, C. (2005). Effects of Test Paper Color on Student Exam Performance: Does It Make a Difference? *Proceedings of the Marketing Management Association Fall Educators' Conference,* 82-87.

Shacklett, J.J., McClure, N. R., & Camiz, J. P. (2005). Security vs. Fairness: The Impact of Colored Paper on Student Test Performance. *Proceedings of American Society of Business and Behavioral Sciences,* 12, (1), 1731-1735.

Titard, P. L., & DeFranceschi, J. E. (2006). The Effect of 'White' vs. 'Colored' Exams on Performance. *Journal of Business and Behavioral Sciences,* 14, (1), 107-113.