Effect of Paper Color on Students' Physics Exam Performances

David R. Schmidt, Todd G. Ruskell, and Patrick B. Kohl

Department of Physics, Colorado School of Mines, 1523 Illinois, Golden, CO 80401

Abstract. Prior work has established the existence of a color-performance relationship in achievement contexts and has demonstrated its presence in some undergraduate course examinations. This study examines the manifestation of such a relationship in an introductory, 430-student, calculus-based electricity and magnetism course during which the paper color used in examinations was varied. In this report, we analyze three separate exams and differentiate between students' multiple choice, written response, conceptual, and computational performances. Also considered are factors such as the time students require to complete exams and their confidence levels prior to and immediately following assessment. Performance in all categories appears to be independent of paper color.

Keywords: Paper Color, Exam Performance
PACS: 01.40.Fk

INTRODUCTION

Research in the cognitive sciences has established that color, acting as an environmental cue, can significantly affect performance within the context of intellectual achievement. Since papers of various colors can be used to distinguish exam versions (as is done at Colorado School of Mines), the possibility that different colors may lead to measurable differences in performance raises an issue of basic fairness in educational practice.

In a multipartite study, Soldat et al. tested red and blue colored paper for conveyed affect and observed subject performance on analytic questions printed thereon [1]. Results showed positive and negative affect (i.e., happiness and sadness) respectively conveyed by red and blue paper and significantly better performance associated with the latter. They argued that color serves as an affective cue which indicates the seriousness of a situation and thereby, the associated degree of processing called for.

In a different study, Elliot et al. found that the perception of red, relative to green or grayscale colors, decreased performance on assessments of mental ability (e.g. IQ, anagrams) and evoked higher levels of avoidance motivation [2]. They attributed this to the association of red with failure in achievement contexts and argued that said association induces avoidance mechanisms which undermine performance [3].

These studies, as many others in this field of research, were performed in clinical settings with volunteer participants. As noted by Sinclair et al., levels of motivation should be higher in an actual examination setting, and it is thus uncertain whether performance effects of color would persist or be overridden [4]. Subsequent studies have yielded inconsistent results. Sinclair et al. observed superior performance on exams printed on blue, as opposed to red, paper whereas others observed contradictory or null results [5-6]. The incongruity of these findings is perplexing given that all of these experiments were conducted with very similar populations (i.e., introductory psychology students) in nearly identical settings (i.e., regular examinations). The few studies conducted on different populations have generally reported null results [7-8].

An explanation for the discrepancies between many prior studies' findings was proposed by Elliot and Maier in 2007 [9]. In a review of prior studies, they found substantial deficiencies in experimental designs and methods. Issues cited emphasized a lack of attention to basic experimental procedure (e.g., blindness), uncontrolled exposure to the colors under scrutiny, and inattention to finer color specifications (e.g., hue, saturation, lightness).

In this study, we examine the manifestation of a color-performance relationship in the field of physics, addressing the issues emphasized by Elliot and Maier in the process. Experimentation was performed during undergraduate course examinations, presumably involving high subject motivation. In contrast to the subject matter of prior studies, physics requires both conceptual understanding and complex mathematical manipulation. Thus, in our analysis, we differentiate between students' conceptual and mathematical performances as well as multiple choice and written response scores. We also analyze students' confidence levels, perceptions of performance, and the amount of time they require in completing exams.

In summary, our study sought to:

(1) Examine the manifestation of a color-performance relationship in an introductory E&M course.
(2) Discuss findings in the context of prior studies' arguments and claims.
(3) Address the implications of results for current institutional practices.

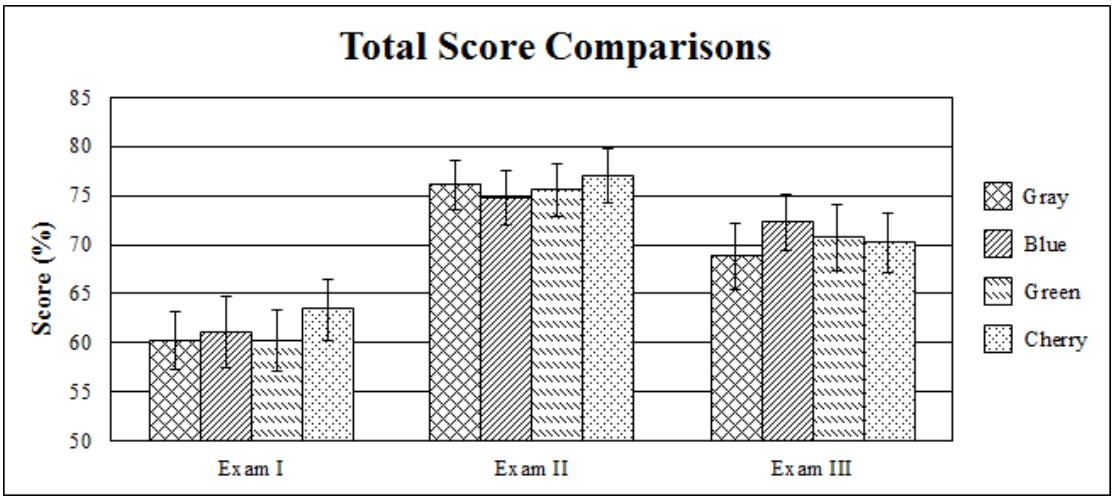

Figure 1. Students' total exam scores differentiated by paper color for Exams I, II, and III. Error bars represent 95% confidence intervals based on standard errors of the means.

METHODS

This study was administered in a calculus-based introductory E&M course at the Colorado School of Mines with an enrollment of 430 students. Excluding the comprehensive final, there were three regular examinations, each accounting for 15% of students' course grades. Given this percentage, students had to perform well on each assessment in order to receive a good mark in the course. Accordingly, most students possessed high levels of motivation and engaged in preparation prior to assessment. This motivation and preparation is in sharp contrast to the situation in studies involving clinical settings [1-2].

Each exam consisted of a multiple choice and free response portion. Multiple choice portions featured both purely conceptual questions and questions involving the manipulation of equations and variables, which in some cases required numerical evaluation to arrive at the correct answer. Free response portions required students to both show and explain each step of their work in order to receive full credit.

The first question of the first and third exams' multiple choice sections asked students to assess their confidence level regarding the exam in its entirety before proceeding. Their options ranged from very low to very high and corresponded to a 1-5 scale. The last question on each of these exams asked students to provide a self-assessment of their performance on the exam as a whole. Their options ranged from very poor to very well and also corresponded to a 1-5 scale.

For each exam, four versions were administered. Each version was printed on a different color paper, the RGB color specifications of which may be seen in Table 1. We chose a diverse set of colors to add breadth to the testing spectrum.

TABLE 1. Paper Color Specifications [10]. Gray corresponds to Wausau Exact Multi-Purpose Gray. Blue and green correspond to Wausau Astrobrights Lunar Blue and Gamma Green respectively. Cherry corresponds to an option of Boise paper whose RGB specifications were unknown. As such, the graphics editor Inkscape was used for estimation.

Name	R	G	B
Gray	212	212	208
Blue	0	187	218
Green	0	154	99
Cherry	255	63	126

Aside from color, exam versions varied only in the order of multiple choice options and the values given for numerical evaluations. Students recorded multiple choice answers on white scantrons and completed free responses on the examination paper itself. Exam administration was performed by teaching assistants who, along with the students, were unaware of the experiment. Said assistants were instructed to record students' completion times but were ignorant as to the reason why.

The same teaching assistants were also responsible for grading the free response portions and were provided with a strict rubric. In order to prevent grader bias due to color, each free response sheet was run through a photocopier with contrast and brightness settings that resulted in uniformly gray outputs. As a secondary measure, the free responses delegated to each teaching assistant included a variety of previously different colored versions.

Important to note is that ~100 students' data was rendered invalid during Exam I due to mismatched paper colors on their multiple choice and free response sections. Teaching assistants in a few exam rooms also failed to record data on the time students required to complete Exams I and III.

TABLE 2. Results from ANOVA testing comparing performances associated with each paper color. The categories examined were students' total score, multiple choice, free response, conceptual, and mathematical performances. Results for Exams I, II, and III are included. Sample sizes range from 314 to 409 students.

	Exam I		**Exam II**		**Exam III**	
	F Ratio	*p-value*	*F Ratio*	*p-value*	*F Ratio*	*p-value*
Total Score	F(3,314)=0.86	p=0.454	F(3,409)=0.47	p=0.702	F(3,395)=0.80	p=0.497
Multiple Choice	F(3,314)=1.21	p=0.306	F(3,409)=0.48	p=0.698	F(3,395)=0.73	p=0.536
Free Response	F(3,314)=0.27	p=0.265	F(3,409)=0.19	p=0.902	F(3,395)=1.53	p=0.206
Conceptual	F(3,314)=0.84	p=0.836	F(3,409)=0.33	p=0.327	F(3,395)=0.85	p=0.849
Mathematical	F(3,314)=0.18	p=0.178	F(3,409)=0.64	p=0.592	F(3,395)=0.43	p=0.731

DATA

Given the ample amount of data obtained from each of the three trials, there are a number of analysis options available. These include comparisons between students' total exam scores, performances on differing question formats, and demonstrated conceptual and mathematical abilities.

Total Exam Scores

Students' total exam scores were a composite of the multiple choice (80%) and free response (20%) portions of an exam and were calculated by dividing the sum of acquired points by the sum of possible points. As our experiments sought to compare the means of multiple groups, we performed one-way ANOVA with a Bonferroni correction. This revealed no significant differences in total exam scores among papers of different color (see Table 2). There were also no distinguishable trends spanning the three exams. Figure 1 provides a visual representation of students' total score performances.

Question Formats and Distinctions

Given the differences associated with multiple choice and free response questions (e.g. partial credit, showing work), it is prudent to analyze each separately for color-based differences in student performance. As before, scores for each format were calculated by dividing acquired points by possible points. We applied one-way ANOVA testing to these data and found no significant differences between colors (see Table 2).

The distinction between conceptual and mathematical problems can be unclear, as many questions blend the two. For the purposes of this study, a question was categorized as "Conceptual" if it did not require any manipulation or numerical evaluation. All other questions were categorized as "Mathematical." On Exam I, the ratio of conceptual to mathematical questions was 11:8. Exam II had a very different ratio of 1:15 whereas Exam III's was similar at 7:9. One-way ANOVA testing yielded no significant performance differences in either of the categories (see Table 2).

Performance Binning and Gains

To test for the possibility that higher and lower performing students are affected differently by color, students' data was divided into three bins on a per-exam basis: those with total exam scores in the lower, middle, and upper third of all students. With these divisions, we performed one-way ANOVA once more. Five such tests yielded p-values below 0.05, but in each case, results were inconsistent across the other two exams. Significance was also diminished when correcting for the number of comparisons drawn.

Individual students' gains were also examined. If students tested on differently colored paper for two separate exams, their score on the first exam was subtracted from that on the second. Students with the same color transitions were grouped together and comparisons were drawn between the different groups. This was done with students' total score data as well as with all of the differentiations discussed above. Results were congruent with those prior; effects were weak at most, and there was no consistent pattern among the various color combinations.

Additional Measurements

Table 3 shows the p-values of one-way ANOVA testing on the time students required to complete their exams, self-reported confidence levels, perceptions of performance, and confidence gains. Confidence gains were calculated through subtraction of students' self-reported levels of confidence from their indications of perceived performance. Differences were mostly insignificant. Two tests yielded p-values under 0.05, but these results were inconsistent across the other two exams in each category. Furthermore, these results would not remain significant when correcting for the number of comparisons drawn. As before, we conducted additional analyses on data binned by total score and on individual student gains; neither approach changed the outcome.

TABLE 3. Results from ANOVA testing comparing non-performance measures associated with each paper color. The categories examined were the time students' required in completing their exams, initial levels of confidence, and perceptions of performance. Results for Exams I, II, and III are included. Sample sizes range from 212 to 393 students. Confidences and perceptions of performance were not measured for Exam II.

	Exam I		**Exam II**		**Exam III**	
	F Ratio	*p-value*	*F Ratio*	*p-value*	*F Ratio*	*p-value*
Time	F(3,212)=2.68	p=0.048	F(3,405)=2.24	p=0.083	F(3,317)=0.41	p=0.745
Initial Confidence	F(3,314)=0.63	p=0.597	X	X	F(3,393)=2.68	p=0.047
Perceived Performance	F(3,303)=1.57	p=0.197	X	X	F(3,382)=1.77	p=0.153
Confidence Gain	F(3,303)=1.13	p=0.339	X	X	F(3,382)=0.56	p=0.640

DISCUSSION AND CONCLUSIONS

The insignificance of all comparisons made by this study is significant as it prompts the question of why previously documented effects of color in achievement contexts were not apparent in our experimental setting.

In the context of Sinclair et al.'s argument, one could argue that the colors used in this study are similar in their indications of situational seriousness and thus, do not incur differences in processing. Our inclusion of starkly contrasting shades of blue, green, and red causes us to doubt the validity of such an argument as the primary cause of our null result.

In the framework of Elliot et al.'s claims, one could argue that the colors used in this study either do not have associations that would hinder or enhance performance, or have similar associations. There are many conflicting claims as to the associations of various shades of blue, green, and red in achievement contexts. As such, we cannot comment on the validity of this argument with regard to our study.

Perhaps the most compelling argument, expressed by Sinclair et al. and others, is that higher levels of motivation could override the potential effects of color. Students in our experiment were engaging in high-stakes examinations based on material for which most had studied, as opposed to low-stakes tasks whose nature is being learned about for the first time during trials.

More experimentation with greater variety therein (e.g., think-aloud interviews) coupled with additional, detailed data analysis would be necessary to make more concrete claims regarding causality. In addition to being academically interesting, our results have immediate implications regarding the use of differently colored paper to distinguish exam versions, which is a common practice at Colorado School of Mines and other institutions. We can say that use of the colors tested here in similar examination settings will most likely not yield differential performances. Depending on the cause of our study's null result, this may or may not be generalizable to all possible colors and shades thereof.

ACKNOWLEDGMENTS

Thanks to the NSF TUES grant 0836937 which funded this study and to the rest of the Physics II teaching staff.

REFERENCES

1. A.S. Soldat, R.C. Sinclair, and M.M. Mark, "Color as an environmental processing cue: External affective cues can directly affect processing strategy without affecting mood," *Soc. Cognition* **15**(1), 55-71, 1997.
2. A.J. Elliot, M.A. Maier, A.C. Moller, R. Friedman, and J. Meinhardt, "Color and psychological functioning: The effect of red on performance attainment," *J. Exp. Psychol. Gen.* **136**(1), 154–168, 2007.
3. A.J. Elliot, "A conceptual history of the achievement goal construct," *Handbook of competence and motivation*, Guilford Press, NY, 52-72, 2005.
4. R.C. Sinclair, A.S. Soldat, and M.M. Mark, "Affective cues and processing strategy: Color-coded examination forms influence performance," *Teach. Psychol.* **25**(2), 130-132, 1998.
5. N.F. Skinner, "Differential test performance from differently colored paper: White paper works best," *Teach. Psychol.*, **31**(2), 111–112, 2004.
6. I.R. Tal, K.G. Akers, and G.K. Hodge, "Effect of paper color and question order on exam performance," *Teach. Psychol*. **35**(1), 26-28, 2008.
7. R. Clary, J. Wandersee, and J.S. Elias, "Does the color-coding of examination versions affect college science students' test performance? Countering claims of bias," *J. Coll. Sci. Teach.* **37**(1), 40–47, 2007.
8. M. Meyer and J.A. Bagwell, "The Non-Impact of Paper Color on Exam Performance," *Issues Account. Educ.* In-Press, 2012.
9. A.J. Elliot and M.A. Maier, "Color and psychological functioning," *Curr. Dir. Psychol. Sci*. **16**(5), 250-254, 2007.
10. Wausau Paper, "Copy of CMYK PMS RGB SWOP Values for Wausau Colors," Spreadsheet, 2011.