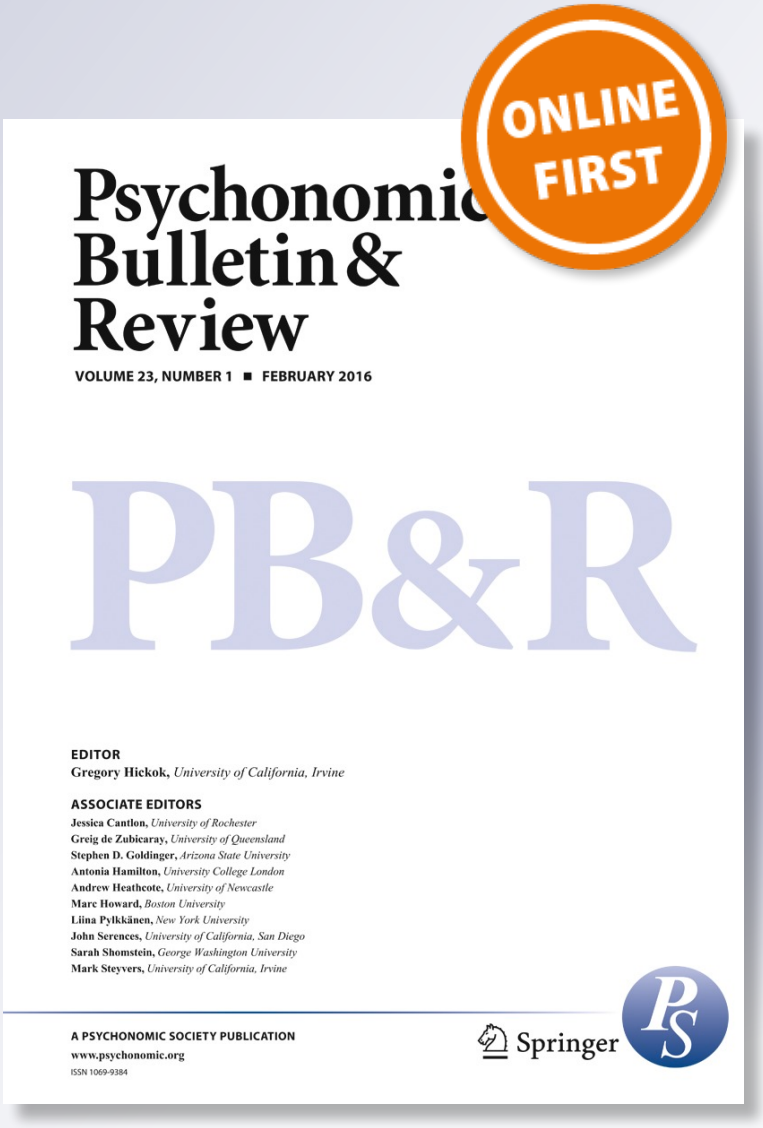
Red vs. green: Does the exam booklet color matter in higher education summative evaluations? Not likely

Winfred Arthur, Inchul Cho & Gonzalo J. Muñoz

Psychonomic Bulletin & Review

ISSN 1069-9384

Psychon Bull Rev
DOI 10.3758/s13423-016-1009-6



Springer

Psychon Bull Rev
DOI 10.3758/s13423-016-1009-6

BRIEF REPORT

Red vs. green: Does the exam booklet color matter in higher education summative evaluations? Not likely

Winfred Arthur Jr.[1] · **Inchul Cho**[1] · **Gonzalo J. Muñoz**[1,2]

Abstract We examined the so-called "red effect" in the context of higher education summative exams under the premise that unlike the conditions or situations where this effect typically has been obtained, the totality of factors, such as higher motivation, familiarity with exam material, and more reliance on domain knowledge that characterize high-stakes testing such as those in operational educational settings, are likely to mitigate any color effects. Using three naturally occurring archival data sets in which students took exams on either red or green exam booklets, the results indicated that booklet color (red vs. green) did not affect exam performance. From a scientific perspective, the results suggest that color effects may be attenuated by factors that characterize high-stakes assessments, and from an applied perspective, they suggest that the choice of red vs. green exam booklets in higher education summative evaluations is likely not a concern.

Keywords Educational testing · Test booklet color · Red · Green · Exam performance

Given the prevalence of color in the human processing of visual information and stimuli, much attention has been paid to color and its relationship with performance in various contexts (Elliot & Maier, 2014; Elliot, Maier, Binser, Friedman & Pekrun, 2009). In particular, research has focused on the extent to which specific colors facilitate or impair performance in achievement contexts. There are several theoretical and conceptual precepts forwarded for the effect of color on performance. For instance, drawing on affect-as-information theory, Soldat, Sinclair, and Mark (1997) argued that color serves as a cue that provides feedback about the nature of a situation and its associated processing requirements. Consequently, exposure to certain colors influences processing strategies much like mood, motivation, or facial expressions, which may facilitate or hinder cognitive performance. Elliot, Moller, Friedman, Maier, and Meinhardt (2007) also posited that color can carry a psychologically positive or negative meaning and that each color is associated with specific actions, such as approach or avoidance behaviors. For instance, according to this model, red is associated with danger or failure, an association that leads to avoidance behavior, which in turn decreases performance in achievement contexts. Elliot and Maier (2012) have since expanded on this general conceptual framework to propose the color-in-context theory, which is "designed to be a broad model of color and psychological functioning that can be used to explain and predict relations between color and affect, cognition, and behavior." (p. 66). However, counter to the theoretical propositions for the relationship between red and performance (Elliot et al., 2007), an *opposite* effect has sometimes been obtained (Mehta & Zhu, 2009; Skinner, 2004), and a number of others have failed to obtain *any* color-related effects in achievement contexts (Clary, Wandersee, & Elias, 2007; Larsson & von Stumm, 2015; Meyer & Bagwell, 2012; Steele, 2014).

A close reading of this body of research indicates that there are noticeable methodological differences between the studies that may account for the mixed findings. Specifically, in addition to being characterized by the use of a very wide range of colors (e.g., blue, green, yellow, red as in Skinner (2004), and

✉ Winfred Arthur, Jr.
w-arthur@tamu.edu

Inchul Cho
iccho83@gmail.com

[1] Department of Psychology, Texas A&M University, College Station, TX, USA

[2] Department of Psychology, Universidad Adolfo Ibáñez, Santiago, Chile

Published online: 16 February 2016

gray, blue, green, and cherry as in Schmidt, Ruskell, and Kohl (2013)), the extant studies differ and covary in terms of test type (e.g., information-processing abilities vs. knowledge-based tests), content assessed (e.g., familiar vs. unfamiliar material), and the outcomes associated with the assessments (i.e., high- vs. low-stakes). Thus, studies that have obtained a color effect appear to be typically conducted under low stakes conditions where participants are tested on unfamiliar material that is ability-based (e.g., anagrams), and the resultant test scores—which are low-stakes because they neither help nor impede the participants' achievement of any of their educational or life objectives—are inconsequential. Studies that are representative of this include Elliot et al. (2007); (2009); Mehta and Zhu (2009), and Soldat et al. (1997). Indeed, all of the studies conducted under these conditions, regardless of whether color is posited to *increase* or *decrease* performance tend to be supportive of color effects on performance. (One exception to the preceding general pattern of results is Steele (2014) which despite having triple the number of participants, and therefore sufficient power to detect the effects, failed to replicate the Mehta and Zhu (2009) color [blue/red] priming effect for the solution of anagrams.)

In contrast, studies conducted under high-stakes conditions which are characterized by participants being tested on knowledge-based content with which they are familiar (e.g., a course final), and the participants have a high-stakes vested interest in the test outcomes (e.g., course grade), often fail to find a color effect (e.g., Clary et al., 2007; Meyer & Bagwell, 2012; Michael & Jones, 1955; Schmidt et al., 2013; Tal, Akers, & Hodge, 2008). (As with the low-stakes studies, one exception to this general pattern of results is Skinner (2004), a high-stakes assessment study that obtained a color effect using a wide range of colors). Although these features are apparent after a detailed review of said studies, the authors do not explicitly acknowledge or recognize and subsequently discuss these characteristics as potential or plausible explanations for their findings. Hence, in the context of higher education summative evaluations, the present study posits that unlike low-stakes research conditions, operational educational assessments comprise familiarity with the exam material, more reliance on domain knowledge, and because the outcomes are high-stakes, a concomitant higher level of motivation. In their totality, we posit that these characteristics are likely to offset any color effects.

Color effects in higher education summative evaluations

As previously alluded to, there is a lack of attention to factors present in high-stakes assessments, such as higher education summative evaluations, that may mitigate color effects. For instance, concerning differences in test/exam content, as noted by Ackerman (2007), research-based tasks (e.g., anagrams) have been stripped of any real-world context. Consequently, information-processing abilities, such as abstract reasoning and working memory are most related to performance on these tasks. In contrast, real-world tasks, such as course exams, are characterized by the fact that they allow individuals to use what they know (i.e., using existing knowledge) in taking the assessment. Thus, unlike the types of tasks that characterize research-based studies of the color effect, operational field-based assessments, such as summative evaluations in higher education, are characterized by familiarity with the exam material and more reliance on domain knowledge.

In addition, because they are high-stakes assessments, motivation is another factor that may attenuate expected color effects. Motivation directs attention and effort and has been consistently shown to have strong, demonstrable effects on purposeful and proactive behavior and performance (Locke & Latham, 2002; Schmidt, Beck, & Gillespie 2013). Thus, consonant with the dynamics of higher education summative evaluations, test-takers are likely to be motivated to achieve high scores and consequently, expend considerable time preparing for exams. In contrast, in the typical low-stakes study, participants usually arrive for the study with relatively low or limited vested interest in performing well because failure to perform well has no important consequences for them. So, the differences in motivation, along with familiarity with exam material and reliance on domain knowledge, are posited as factors that in their totality differentiate research-based tasks from the sorts of exams that characterize higher education summative evaluations; and collectively, are likely to result in a mitigation of the color effects that have been observed in low-stakes conditions.

Study overview

There are several reasons why an examination of the effect of the color of test booklets on exam performance in higher education summative evaluations is important. First, from a practical perspective, there are several reasons why course instructors, test administrators, and other evaluators may use different color test booklets in large scale testing settings (e.g., distinguish the alternative forms if different versions of the exam are being used in the same administration, and also to mitigate or try to prevent cheating). Second, as is forwarded here, it is plausible and indeed likely that color effects observed under low-stakes conditions may not generalize to higher education summative evaluations and other high-stakes testing situations. Several researchers (Ferguson, 2004; Mitchell, 2012) have expressed concern about the generalizability of findings across settings; and whereas they specifically speak to lab vs. field settings, in the present instance our focus is on the generalizability across low- and high-

stakes testing conditions. Hence, although studies such as Elliot et al. (2007) have been conducted under highly controlled experimental conditions with the associated gains in internal validity, summative evaluations in higher education typically involve situations in which test-takers are motivated to perform well and also are tested on content for which they have had the opportunity to prepare or are otherwise familiar with the material. In addition, Elliot et al. (2007) appear to acknowledge this potential generalizability threat when they note that "our results were obtained using brief, controlled color presentations placed directly on the achievement task, and the generality of the effect beyond these parameters needs to be tested" (p. 165). In addition, Elliot and Maier (2012) also note that the failure to consider color-in-context "is largely responsible for the accumulation of inconsistent empirical data that have hampered progress and growth in this promising area." (p. 72). Consequently, building on the conceptual framework of Elliot and colleagues, the present study sought to examine the effect of exam booklet color, specifically red vs. green, on summative evaluations in higher education.

The research objectives were accomplished by opportunistically capitalizing on three naturally occurring archival data sets in a higher education setting. Whereas we had access to several other colored exam data sets (e.g., yellow, blue, purple, etc.), we chose to use only the red and green data sets in the present study because these colors have been the focus of Elliot and colleagues with green serving as the chromatic contrast for red. According to Elliot et al. (2007), green is associated with approach-oriented behaviors (e.g., "go" in traffic lights), whereas red signifies avoidance-oriented behaviors (e.g., "stop" in traffic lights). So, in one data set (henceforth referred to as Study 1), students in an undergraduate psychology research methods class took a final exam in which *only the cover page* of their exam booklet was colored, red or green. In another data set (Study 2) the *entire exam booklet* was colored, again red or green and thus was considered to be a stronger test than Study 1 and an extension of previous research in which only the color of the cover page was manipulated to test the red effect in a high-stakes condition (Smajic, Merritt, Banister, & Blinebry, 2014). Finally, whereas the exams used in Study 1 and 2 contributed 21 % and 15 % respectively to the students' final course grade, in the third data set (Study 3), the colored exam was a baseline exam (administered on the first day of class for diagnostic purposes) that did not contribute to the students' course grade; but it was similar to Study 2 in that the entire exam booklet was on colored paper.

Method

Participants

The participants for all three studies were students enrolled in an undergraduate psychology research methods course in a large public U.S. university. Because these were nonintrusive naturally occurring data sets, no sex or age data were available for the samples. However, the samples were typical of the population taking this type of course, namely predominately female and approximately 19–20 years old. There were 76, 164, and 87 students in Studies 1, 2, and 3, respectively.

Measures

The measure used to operationalize the dependent variable was a multiple-choice knowledge exam. Concerning the independent variable, for Study 1 *only the cover (double-sided) page* was on red or green paper; the rest of the exam was on white paper. For Study 2, the entire exam booklet was on colored paper. For Studies 1 and 2, the total exam score contributed 21 % and 15 %, respectively, to the students' course grade. We describe Study 3 as low-stakes, because the exam was a baseline exam that did not contribute to the students' course grade. However, its content was basic statistics, covering material from a prior prerequisite course that is typically taken in the preceding semester. Lastly, like Study 2, in Study 3 the entire booklet was on colored paper. Table 1 presents a summary of the exams and other study characteristics.

Procedure

All exams were administered during the scheduled class period. All exam booklets were randomly distributed; thus for colored exams, randomly assigning students to the red and green conditions.

Results

Although students randomly received red and green exams, because these were naturally occurring groups, to control for potential preexisting differences between groups, all analyses utilized an analysis of covariance (ANCOVA). For Study 1, the covariates utilized were a baseline performance test on white paper, the items on the comprehensive final repeated from three previous exams administered during the semester on white paper, and GPA. For Study 2, a baseline test on white paper and GPA were used as covariates, whereas for Study 3 only GPA was used as a covariate. Results for Studies 1, 2, and 3 are presented in Table 2. For Study 1, the results indicated that despite a condition that was stronger than Elliot et al.'s (2007) manipulation of having only the participants' identification numbers in color, there were no performance differences between the color conditions, $F(1, 65) = 2.11$, $p > 0.05$, $\eta^2 = 0.03$. Consistent with the results from Study 1, Study 2 provided further evidence that exam performance was not affected by booklet color even when the entire test booklet was printed on red or green paper $F(1, 154) = 0.30$, $p > 0.05$,

Table 1 Summary of exams and study characteristics

	Study 1	Study 2	Study 3
No. of participants	76 (G = 40, R = 36)	164 (G = 83, R = 81)	87 (G = 44, R = 43)
Color condition	cover page only	entire exam booklet	entire exam booklet
No. of items			
Total items	92	96	17
Items on colored pages	9 (of 92)	96	17
Dependent variable	% correct of 9 items on colored pages	% correct of 96 items	% correct of 17 items
Covariates	GPA	GPA	GPA
	Baseline (17 items)/white	Baseline (17 items)/white	
	"Pretest" total; same 92 items as comprehensive final repeated from previous 3 exams administered during the semester/white		
Purpose of exam	high-stakes → comprehensive final	high-stakes → exam 1 (of 4)	low-stakes → diagnostic baseline exam
Percentage of course grade	21%[a]	15 %	0 %
When administered	end of semester	4 weeks into semester	first day of class

G = green condition, R = red condition, GPA = grade point average (self-reported on baseline exam).

[a] Total exam, not just the 9 items on colored paper.

$\eta^2 = 0.00$. Finally, as with Study 1 and 2, the results for Study 3 indicated that the green and the red groups did not differ on a baseline exam that was administered on the first day of the course, $F(1, 80) = 0.18$, $p > 0.05$, $\eta^2 = 0.00$. Therefore,

Table 2 Descriptive statistics and performance differences between red and green exam booklet conditions

Variables	Green condition		Red condition			
	M	*SD*	*M*	*SD*	*d*	η^2
Study 1	*n* = 40		*n* = 36			
Covariates						
Baseline/white	52.31	10.72	56.40	11.23	−0.37	
GPA	3.05	0.48	3.16	0.43	−0.24	
"Pretest" total/white[a]	73.77	9.56	77.91	7.20	−0.48*	
Color condition						
Colored cover (double-sided) page only	80.83	14.23	86.11	9.34	−0.43	.03
Study 2	*n* = 83		*n* = 81			
Covariates						
Baseline/white	53.86	15.97	55.38	13.70	−0.10	
GPA	3.14	0.42	3.17	0.41	−0.07	
Color condition						
Colored comprehensive final exam	74.89	10.19	75.13	11.12	−0.02	.00
Study 3	*n* = 44		*n* = 43			
Covariates						
GPA	3.15	0.42	3.17	0.43	−0.05	
Color condition						
Colored baseline	48.93	12.92	49.93	10.45	−0.09	.00

[a] Same 92 items as comprehensive final repeated from previous 3 exams administered during the semester/white.

GPA = grade point average.

η^2 is based on ANCOVA controlling for the specified covariates. All exam scores were converted to percentages.

*$p < 0.05$ (two-tailed).

consistent with the results for Study 1 and 2, Study 3 provided further evidence that exam performance was not affected by booklet cover even under conditions of weaker familiarity and motivation.

General discussion

Elliot and Maier (2012) present a compelling case for the "importance of attending to color as a psychologically relevant stimulus" (p. 109) in the domains of achievement and affiliation. With a focus on achievement in educational and employment settings where tests and other assessment measures may be printed on colored paper or presented on colored computer screens, the choice of color may be an important consideration. Consequently, building on Elliot and colleagues' work, the present study examined the extent to which red vs. green exam booklets influences summative evaluation scores in a college course. Like Elliot and colleagues, we focused on red with green as its chromatic contrast, because like red, green "is an additive primary color and is considered the opposite of red in several well-established color models" (Elliot & Maier, 2012, p. 77). In terms of our findings, unlike previous results from studies conducted under low-stakes conditions, the effect of red vs. green exam booklets on performance was negligible at best, despite the relatively high levels of power to detect said effects.[1] Thus, neither the exam cover-only nor stronger whole-exam color conditions yielded the effects that have been observed in low-stakes conditions. This was even the case with Study 3 for which, although the exam scores did not contribute to students' course grades (because it was a baseline exam), it nevertheless covered material that the students had just covered in the previous semester. Although this was a low-stakes exam, these latter results suggest that testing students on familiar content may be sufficient to offset the color effects observed with ability-based tests that utilize mostly unfamiliar or novel materials.

Nevertheless, the results of the present study can be interpreted as being consistent with the tenets of Elliot and Maier's (2012) color-in-context theory in that our hypotheses were based on the premise that the high stakes and thus higher motivation, familiarity with exam material, and reliance on domain knowledge would in their totality offset the avoidance motivation effects posited by Elliot and colleagues to account for the negative effects of red on achievement on intellectual performance. Hence, our results suggest that context matters—that when participants are tested on a high-stakes knowledge exam for which they have prepared, taking said exams on red (vs. green) test booklets is unlikely to affect their exam performance. Consequently, the failure to take this boundary condition into account maybe an important contributory factor to the inconsistency in empirical findings in this domain.

Limitations and future directions

Color stimuli vary on hue, lightness, and chroma, and Elliot and Maier (2012) have observed that nearly all the extant literature has failed to attend to this in the manipulations of color. They also recommend two approaches of either selecting hues on the basis of color samples with a well-validated color model or using a spectrophotometer to assess color at the spectral level. A limitation of the present study is that it did not use either of these approaches in selecting and matching the red and green hues used on lightness and chroma. This is because these were naturally occurring data sets that were not designed to exam the issues addressed; however, they opportunistically allowed us to do so because we have been using various color exams in our courses for test security and other test administration purposes. Consequently, given the applied nature of the present test, this limitation was not considered to be a particularly onerous threat since from an ecological validity perspective, the choice of red vs. green colors was representative of what the typical course instructor or test developer would do in selecting colors for their exam booklets or computer screens. Also, if color effects surface under such unique conditions, then the impact of this specific test design feature on student performance also will be reduced considerably. Nevertheless, it is a limitation that serves as an avenue for future research in replicating the present results. In addition, as demonstrated in other studies (Schmidt et al., 2013; Skinner, 2004), educators and other assessment practitioners may use a wider range of colors than just red and green; our study was limited to a comparison of red vs. green.

We also acknowledge that we conceptualized high-stakes conditions as being collectively characterized by factors such as the use knowledge-based tests, familiar content material, and valued outcomes. However, our study did not separately examine the effects of these characteristics and so we are unable to speak to their separate contributory effects; our results only pertain to the totality of their effects.

Elliot and colleagues have demonstrated the effects of red on not only performance but on avoidance behaviors as well. As such, avoidance behaviors associated with red might slow the response time of test takers (Elliot et al., 2009). Therefore, it is plausible that red vs. green test color effects may be more critical with speeded tests compared with power-based tests like those used in the present studies. Hence, this serves as another possible avenue for future research.

Finally, whereas our exams were administered on paper, admittedly, there is a continued increase in computer-administered tests. Consequently, future research could

[1] For example, given our smallest sample size [$N = 76$ for Study 1] and largest sample size [$N = 164$ for Study 2], our power to detect the effect sizes in the four studies reported in Elliot et al. [2007, $\eta^2 = 0.09$ to 0.50] ranged from 0.76 to 1.00.

examine the extent to which the effects or lack thereof are replicated with colored computer screens. One would expect them to be replicated, but it is an empirical issue. In conclusion, within the context of the limitations noted above, our results suggest that the choice of red vs. green exam booklets in higher education summative evaluations is likely not a concern.

References

Ackerman, P. L. (2007). New developments in understanding skilled performance. *Current Directions in Psychological Science, 16,* 235–239.

Clary, R., Wandersee, J., & Elias, J. S. (2007). Does the color-coding of examination versions affect college science students' test performance? Countering claims of bias. *Journal of College Science Teaching, 37,* 40–47.

Elliot, A. J., & Maier, M. A. (2012). Context-in-color theory. *Advances in Experimental Social Psychology, 45,* 61–125. doi:10.1016/B978-0-12-394286-9.00002-0

Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning. *Annual Review of Psychology, 65,* 95–120. doi: 10.11146/annurev-psych-010213-115035

Elliot, A. J., Maier, M. A., Binser, M. J., Friedman, R., & Pekrun, R. (2009). The effect of red on avoidance behavior in achievement contexts. *Personality and Social Psychology Bulletin, 35,* 365–375. doi:10.1177/0146167208328330

Elliot, A., Moller, A., Friedman, R., Maier, M., & Meinhardt, J. (2007). Color and psychological functioning: The effect of red on performance attainment. *Journal of Experimental Psychology: General, 136,* 154–168. doi:10.1037/0096-3445.136.1.154

Ferguson, L. L. (2004). External validity, generalizability, and knowledge utilization. *Journal of Nursing Scholarship, 36,* 16–22. doi:10.1111/j.1547-5069.2004.04006.x

Larsson, E. E. C., & von Stumm, S. (2015). Seeing red? The effect of colour on intelligence test performance. *Intelligence, 48,* 133–136. doi:10.1016/j.intell.2014.11.007

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57,* 705–717. doi:10.1037//0003-066X.57.9.705

Mehta, R., & Zhu, R. (2009). Blue or red? Exploring the effect of color on cognitive task performances. *Science, 323,* 1226–1229. doi:10.1126/science.1169144

Meyer, M. J., & Bagwell, J. (2012). The non-impact of paper color on exam performance. *Issues in Accounting Education, 27,* 691–706. doi:10.2308/iace-50142

Michael, W., & Jones, R. (1955). The influence of color of paper upon scores earned on objective achievement examination. *Journal of Applied Psychology, 39,* 447–450. doi:10.1037/h0045289

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7,* 109–117. doi:10.1177/1745691611432343

Schmidt, A. M., Beck, J. W., & Gillespie, J. Z. (2013a). Motivation. In N. W. Schmitt, S. Highhouse, & I. Weiner (Eds.), *Handbook of Psychology, Volume 12: Industrial and Organizational Psychology* (2nd ed., pp. 311–340). Hoboken, NJ: John Wiley and Sons.

Schmidt, D. R., Ruskell, T. G., & Kohl, P. B. (2013b). Effect of paper color on students' physics exam performances. *Physics Education Research Conference. AIP Conference Proceedings*. doi:10.1063/1.4789730

Skinner, N. F. (2004). Differential test performance from differently colored paper: White paper works best. *Teaching of Psychology, 31,* 111–113. doi:10.1207/s15328023top3102_6

Smajic, A., Merritt, S., Banister, C., & Blinebry, A. (2014). The red effect, anxiety, and exam performance: A multidisciplinary examination. *Teaching of Psychology, 41,* 37–43. doi:10.1177/0098628313514176

Soldat, A. S., Sinclair, R. C., & Mark, M. M. (1997). Color as an environmental processing cue: External affective cues can directly affect processing strategy without affecting mood. *Social Cognition, 15,* 55–71. doi:10.1521/soco.1997.15.1.55

Steele, K. M. (2014). Failure to replicate the Mehta and Zhu (2009) color effect. *Psychonomic Bulletin & Review, 21,* 771–776. doi:10.3758/s13423-013-0548-3

Tal, I., Akers, K., & Hodge, G. (2008). Effect of paper color and question order on exam performance. *Teaching of Psychology, 35,* 26–28. doi:10.1080/00986280701818482