Journal of Personality and Social Psychology
2009, Vol. 96, No. 1, 135–154

0022-3514/09/$12.00 DOI: 10.1037/a0012878

Testing the Mere Effort Account of the Evaluation–Performance Relationship

Sametria R. McFall
Savannah State University

Jeremy P. Jamieson and Stephen G. Harkins
Northeastern University

Research traditions in psychology in which the evaluation–performance relationship was examined do not show agreement on the mediating process, nor is there any compelling evidence that favors one account over the others. On the basis of a molecular analysis of performance on the Remote Associates Test (RAT), Harkins (2006) argued that the potential for evaluation motivates participants to perform well, which potentiates prepotent responses. If the prepotent response is correct, performance is facilitated. If the prepotent response is incorrect, and participants do not know, or if they lack the knowledge or time required for correction, performance is debilitated. The present research pits this mere effort account against 4 other potential explanations (withdrawal of effort, processing interference, focus of attention, and drive) on 3 tasks that were specifically selected for this purpose (anagram solution, the Stroop Color–Word task, and the antisaccade task). In each case, the results are consistent with the mere effort account.

Keywords: mere effort, evaluation, motivation

The effect of the potential for evaluation on task performance has been a topic of interest in social psychology for more than a century (Triplett, 1898). In five different research traditions in psychology, it has been found that the potential for evaluation tends to facilitate performance on simple tasks but to debilitate it on complex ones: social loafing (Jackson & Williams, 1985), goal setting (Locke & Latham, 1990), creativity (Amabile, 1979), achievement goal theory (Elliot, Shell, Henry, & Maier, 2005), and social facilitation (Geen, 1989).

Within these traditions, process models have been proposed to account for these findings, but a review reveals no agreement across, or even within, these traditions (Harkins, 2001). Specifying the mediating process is the key to the theoretical development of each of these research traditions as well as to any effort to integrate them. In addition, when it comes to application, it is impossible to suggest interventions if one does not understand the mediating process. For example, the intervention that would be designed if people were withdrawing effort and failing as a result would be completely different from the intervention that would be proposed if people were trying hard but their efforts were misdirected, leading to failure.

Harkins (2006) argued that the field's failure to resolve this issue may be a result of the fact that our efforts have been focused broadly on theory construction rather than on the tedious analysis required to learn how performance unfolds on a given task. He suggested that although it would appear that a molecular analysis of task performance would be an integral part of theory development, this type of analysis has not been conducted, and that it is possible that the mediating process could be identified through such an approach. To this end, Harkins (2006) undertook a molecular analysis of the effects of evaluation on the performance of a specific task, the Remote Associates Test (RAT).

The RAT requires participants to look at a set of three words (e.g., *lapse, elephant,* and *vivid*) and generate a fourth word that is related to each word in the given triad (in this case *memory*). Harkins (2001) has shown that the potential for evaluation produces the typical pattern of performance on this task: Participants who anticipate evaluation by the experimenter solve more triads shown by a pretest to be simple than do no-evaluation participants, whereas participants who anticipate experimenter evaluation solve fewer triads shown by a pretest to be difficult than do no-evaluation participants.

Harkins's (2006) analysis suggested the mere effort account. In this explanation, he argued that the potential for evaluation motivates participants to want to do well, which potentiates whatever response is prepotent on the given task. On the RAT, the prepotent response is to generate words that are closely related to one of the triad members. Because on simple items the correct answers tend to be a close associate of at least one of the triad members, the greater effort on the part of participants subject to evaluation leads to the production of more close associates and to better performance.

On the other hand, on the complex items, the associations between the triad members and the correct answer are much weaker (i.e., the associates are more remote), and the participants are extremely unlikely to produce the solution by generating associates for the individual triad members. So, for example, if

Sametria R. McFall, Department of Continuing Education, Savannah State University; Jeremy P. Jamieson and Stephen G. Harkins, Department of Psychology, Northeastern University.

Thanks go to Neal Pearlmutter for the use of the eye tracker (support by National Institutes of Health Grant R01-DC05237). Thanks also go to Sean Allen for writing the computer program used in Experiment 1.

Correspondence concerning this article should be addressed to Stephen G. Harkins, Department of Psychology, 125 Nightingale, Northeastern University, 360 Huntington Avenue, Boston, MA 02115. E-mail: s.harkins@neu.edu

presented with the triad member *note*, a participant would be extremely unlikely to produce the associate, "bank." Nonetheless, when the participant considers the word *note*, the solution, "bank," is weakly activated. Likewise, the solution, "bank" is also weakly activated when the participant considers the other two triad members, *river* and *blood*. If this were the only operative process, this weak activation should accumulate over time, leading to the emergence of the correct answer. However, when participants actively test close associates as solutions for the triads, these associates are highly activated, and they strongly inhibit the activation of the remote (weak) associates. Thus, generating close associates, the same behavior that facilitates the performance of participants subject to experimenter evaluation on simple items, debilitates that performance on complex items.

Drive theory (Zajonc, 1965), like the mere effort account, accords a central role to prepotent or dominant responses. In drive theory, it is contended that the presence of others produces arousal, which increases drive. Increased drive enhances the probability of the emission of dominant responses, which are likely to be correct on simple tasks but incorrect on difficult ones. In fact, Cottrell (1968, 1972) argued that this drive was the result of the participants' apprehension about the fact that they could be evaluated.

Thus, both mere effort and Cottrell's (1968, 1972) evaluation apprehension accounts of social facilitation effects predict that the potential for evaluation will potentiate dominant or prepotent responses. However, in the case of mere effort, this potentiation results from the motivation to perform well, which should also lead to an effort to correct an incorrect response if the participant recognizes that his or her response is incorrect, knows the correct response, and has the opportunity to make it. In contrast, Cottrell's (1968) modification of Zajonc's (1965) drive theory suggests only that the positive or negative anticipations produced by the presence of others nonselectively energize individual performance (i.e., potentiate the dominant response). On a task like the RAT, one is unable to distinguish between mere effort and evaluation apprehension accounts because even if the participants know that the response is incorrect, they do not know how to correct it. As a result, one cannot see the effect of the motivation to correct on this task.

At least three other explanations have been proposed to account for the fact that the potential for evaluation debilitates performance on complex tasks: (a) Concern about failure leads to withdrawal of effort (social facilitation: Carver & Scheier, 1981; achievement goal theory: Elliot et al., 2005; social loafing: Harkins, 2001; creativity: Hennessey, 2001); (b) concern about failure diminishes processing capacity (social facilitation: Bond, 1982; achievement goal theory: Elliot et al., 2005; cf. Sarason, Pierce, & Sarason, 1996); and (c) attentional overload restricts focus of attention leading to poor performance on complex tasks, which often require use of a wider range of cues than simple tasks (social facilitation: Baron, 1986; creativity: Hennessey, 2001; Huguet, Galvaing, Monteil & Dumas, 1999; Muller & Butera, 2007).

Harkins's (2006) findings suggest that on the complex RAT items, participants subject to evaluation do not perform poorly because they withdraw effort (Carver & Scheier, 1981; Elliot et al., 2005; Harkins, 2001; Hennessey, 2001). It is the fact that they are putting out effort that is the source of their difficulty on complex triads. It is not that worry concerning failure takes up processing capacity, ensuring failure (Bond, 1982; Sarason, Pierce, & Sarason, 1996). Once again, Harkins's (2006) findings suggest that participants subject to experimenter evaluation are engaged in the same behavior on both simple and complex items. It is just that this behavior is effective on simple items but is ineffective on complex ones.

A third explanation, focus of attention, suggests that the potential for evaluation produces an attentional overload that "leads to a restriction in cognitive focus in which the individual attends more to cues that are most central to the task (or alternatively most central geographically in the display) at the expense of more peripheral cues" (Baron, 1986, p. 27). This cognitive explanation does not account for the role that motivation plays in producing the pattern of results on the RAT. That is, participants who are subject to evaluation do not perform better on simple items because the answer candidates that they generate are more closely related to the triad members (central cues) than are the answer candidates generated by the no-evaluation participants. The participants in the two conditions are equally likely to think of answer candidates that are closely related to the triad members. It is the fact that participants subject to evaluation are motivated to generate and test more of these closely related candidates that accounts for the fact that they outperform no-evaluation participants. On the complex items, it is this same motivation to test more closely related candidates that inhibits the accumulation of the activation required for the correct answer to emerge. Thus, no-evaluation participants do not perform better on complex items than do participants who are subject to evaluation because no-evaluation participants are better able to think of more remotely associated answer candidates (i.e., peripheral cues) than are participants subject to evaluation. No-evaluation participants perform better because the same lack of motivation that prevents them from testing enough closely related candidates to come up with correct answers on simple items allows the small amount of activation produced by each triad member to accumulate to the point that the correct answer "pops out" on the complex items.

Although in the preceding we argue that Harkins's (2006) findings are inconsistent with these three accounts, his research was not designed to test one account against another. In addition, Harkins's (2006) work does not distinguish between the drive/evaluation apprehension and mere effort accounts. In the current research, we used tasks that were specifically selected to pit mere effort against these other accounts. In one experiment, we used an anagram solution task to pit the mere effort explanation against the processing interference account (e.g., Bond, 1982; Elliot et al., 2005) and the withdrawal of effort explanation (e.g., Carver and Scheier, 1981; Harkins, 2001; Hennessey, 2001). In two experiments, we used the Stroop Color–Word Task to test the mere effort explanation against the focus of attention (e.g., Baron, 1986; Huguet et al., 1999) and the drive/evaluation apprehension (Cottrell, 1972; Zajonc, 1965) accounts. And finally, in a fourth experiment, we used the antisaccade task (Hallet, 1978) to pit mere effort against Muller and Butera's (2007) extension of Huguet et al.'s (1999) focus of attention account. In this experiment, we were also able to use an eye tracker, which makes it possible to isolate various components of performance (e.g., prepotent responding, saccade launch latencies, response time measured from when the target can be seen), allowing an assessment of their relative contribution to the terminal measures on this task (identification of target

orientation, and overall reaction time for response). Most tasks do not permit this level of analysis.

Anagrams

Research on anagram solution shows that participants attempt to solve these problems by first trying different letters in the first position, and because many more words begin with consonants than with vowels, participants have a strong tendency to begin with consonants (Witte & Freund, 2001). The mere effort hypothesis predicts that this prepotent response will be enhanced when participants are subject to evaluation by the experimenter. That is, just as participants subject to evaluation are highly motivated to solve the RAT items and, to this end, generate close associates of the triad members, participants subject to evaluation should be highly motivated to solve anagrams and should attempt to do so by testing consonants in the first position. As a result, participants subject to experimenter evaluation should be more likely to solve anagrams of words that begin with consonants but less likely to solve anagrams of words that begin with vowels than are participants who are not subject to evaluation.

Other variables have also been shown to affect the solvability of anagrams. For example, the greater the frequency of the appearance of a word in the language, the easier it is to solve its anagram (e.g., Mayzner & Tresselt, 1958; Witte & Freund, 2001). However, this effect emerges from the tendency to try letters in the first position. Words of high frequency are more easily retrieved from the lexicon than are low frequency words because high frequency words have been encountered more often and, as a result, have a higher level of resting activation than do the words encountered less frequently. This activation provides these words with a head start when they receive additional activation. However, simply looking at a set of five scrambled letters does not do anything to increase the activation of the high frequency words over low frequency words. Current activation-based models of the lexicon (e.g., Plaut, McClellan, Seidenberg & Patterson, 1996) require that the letters in the word appear in the correct position for activation for that answer candidate to be strongly supported. Thus, for word frequency to have its effect in the anagram task, the participant engages the system by placing a letter, most likely a consonant, in the first position. If the candidate word is correct, as the participant then tries other letters in the other positions, the word gains additional activation until it pops out. Or if no correct answer pops out, the participant tries another letter in the first position.

Thus, we argue that the prepotent response is to try consonants in the first position, and the potential for evaluation should potentiate this response. The activation system takes care of the rest. As a result, the mere effort explanation would not predict an interaction between evaluation potential and word frequency. Whether the solutions to anagrams are words of high frequency or low frequency, solvers will still tend to try consonants in the first position, and this prepotent response should be even more likely to be made by participants subject to evaluation than by those who are not, yielding only a main effect for word frequency.

In contrast, the withdrawal of effort and the processing interference accounts would make the interaction prediction for evaluation and word frequency as well as for evaluation and initial letter. That is, words that are high frequency should be experienced as easier to solve than are words that are low frequency, just as words that begin with consonants should be experienced as easier to solve are than words that begin with vowels. Both the withdrawal of effort and processing accounts suggest that participants monitor how well they are performing on a given task. When this self-monitoring indicates that success is not assured, in the processing interference account, it is proposed that worry takes up processing capacity, preventing the correct response from emerging, whereas in the withdrawal of effort account, it is argued that when participants believe that they cannot bring their behavior in line with the standard, they stop trying. According to these accounts, the combination of the potential for evaluation and the experience of the task as difficult should matter, not the source of the difficulty. Thus, it should not matter whether the solution is difficult because the word appears infrequently in the language or begins with a vowel instead of a consonant. In each of these instances, the solver should experience difficulty in achieving success and performance should suffer.

The focus of attention explanation does not make strong predictions for anagram performance. To solve anagrams, the participant must use all the letters to produce a word. Thus, each letter is a central cue in this task. Additionally, the letters are the only stimuli presented on screen and are located in the center of the display. Thus, the notion of centrality, whether cognitive or geographical, would not appear to be an issue on this task. And finally, on the anagram task, as on the RAT, we are unable to distinguish between mere effort and drive/evaluation apprehension accounts because participants are unlikely to know that their prepotent response is incorrect.

Experiment 1: Word Frequency and Initial Letter

In Experiment 1, we manipulated evaluation potential in combination with word frequency and whether the initial letter was a vowel or a consonant. The mere effort and drive/evaluation apprehension accounts predict an interaction between evaluation potential and whether the word begins with a consonant or a vowel. The prepotent response is to test consonants as the initial letter, which should debilitate the performance of participants subject to evaluation on words that begin with vowels but which should facilitate it on words that begin with consonants. However, given the fact that the word frequency effects emerge as a result of the action of the activation system, these accounts would predict a main effect for word frequency but no interaction with evaluation potential. In contrast, the withdrawal of effort and the processing interference explanations would predict that working with low frequency words or working with words that begin with vowels would each produce the experience of difficulty. As a result, these accounts would predict two-way interactions between experimenter evaluation and each of the other variables: word frequency and whether the word begins with a vowel or a consonant.

However, in testing these accounts, we must also consider whether these effects would be expected to emerge at the level of the item or in the aggregate. For example, Bond's (1982) processing interference account predicted "task-wide social effects—effects that depend on aggregate task difficulty" (p. 1044). Thus, in this account, it would be argued that it is the cumulative experience

of success or failure that produces facilitation or debilitation.[1] Other researchers who argued that concern about failure diminishes processing capacity (e.g., Elliot et al., 2005) could make a similar argument, as could researchers who argued that concern about failure leads to withdrawal of effort (e.g., Carver & Scheier, 1981; Elliot et al., 2005; Harkins, 2001; Hennessey, 2001). In contrast, although other manipulations could moderate the effect, given the evaluation manipulation alone, mere effort makes its predictions at the level of the item, as does drive theory.

To test the item account versus aggregate account, we ran two versions of this experiment. In both versions, the evaluation manipulation was a between-subjects factor. In Version 1, initial letter (consonant vs. vowel) was a within subjects factor, whereas word frequency was a between-subjects factor. In Version 2, word frequency was a within subjects factor whereas initial letter was a between-subjects factor. Thus, in the aggregate, when frequency was the between-subjects variable, participants in the low frequency condition should be running into the difficulty that Bond (1982) suggested would lead to processing difficulty (or to withdrawal of effort) and, thus, to performance debilitation, whereas participants in the high frequency condition should be experiencing success in the aggregate. So, these accounts would predict an evaluation by frequency interaction. And likewise, when consonant–vowel is the between-subjects variable, these accounts would predict an interaction between consonant–vowel and evaluation. If researchers other than Bond (1982) wanted to argue for item-level difficulty effects, of course, they would predict two-way interactions between evaluation and word frequency and between evaluation and initial letter, whether these variables were between subjects or within subjects.

In contrast, the mere effort and drive accounts predict that evaluation potentiates the prepotent response, putting consonants in the first position, and so, evaluation should interact with consonant–vowel whether it is a within-subjects variable or a between-subjects variable. These accounts would also predict that there would be a main effect for frequency, whether it is a within-subjects variable or a between-subjects variable. There is no reason that the activation process that produces this effect should be impacted by this manipulation.

Method

Participants

Ninety-six Northeastern University undergraduates (52% female, 48% male) participated in this experiment as a means of satisfying a course requirement. In Version 1 of the experiment, 47 participants were randomly assigned to one of four conditions produced by crossing two between-subjects factors (experimenter evaluation vs. no experimenter evaluation, and high frequency words vs. low frequency words). The third factor, initial letter (consonant vs. vowel), was a within subjects factor. In Version 2, 49 participants were randomly assigned to one of four conditions produced by crossing experimenter evaluation and initial letter, with word frequency serving as a within subjects variable.

Procedure

Participants were run individually on a computer. Upon entering the lab, participants were asked to read instructions that explained that they were going to work on an anagram task in which they were to try to rearrange the scrambled letters to produce a word. They were informed that the letters would form only one word and no proper nouns. Participants were then instructed that they would be given a series of anagrams, each of which they would have 1 min to solve. Participants were shown two example anagrams, one for which the solution began with a consonant and one for which the solution began with a vowel.

After the example anagrams, half of the participants were told that we were interested in their performance as individuals and that we would examine their performance at the end of the session (experimenter evaluation) but that we would be unable to provide individual feedback as this may affect the performance of later participants. The other half of the participants were told that we were interested in average performance and that we would not look at their individual performance (no experimenter evaluation). Instead, they were asked to click a button on the computer labeled *average scores* to score their performance and average it with the performance of previous participants. The participants were given 1 min to solve each of 20 anagrams. Participants were instructed to click the *enter answer* button on the computer when they were ready to solve the anagram; the response box would appear, and they would have 3 s to type their answers. The experimenter then left the room and the participants saw the 20 anagrams.

The experiment was run in two versions. In each version, the evaluation manipulation was a between-subjects factor. Again, in Version 1, initial letter (consonant vs. vowel) was a within-subjects factor (10 words began with a consonant, 10 words began with a vowel), whereas word frequency was a between-subjects factor. In Version 2, word frequency was a within-subjects factor

[1] Bond's (1982) research was aimed at showing that it was the aggregate task experience that determined the task outcome, rather than the item level of performance predicted by drive theory. Thus, he showed that the debilitation found on complex items was eliminated when these items were embedded in an easy task and that performance on simple items suffered when they were embedded in a complex task. However, we argue that his results are compatible with the mere effort account. In his experiment, participants first learned a 13-item paired-associate list to criterion and then were asked to learn a new set of paired-associates. On the easy version of the new list, the response terms for 3 pairs were interchanged, whereas the other 10 were left unchanged. On the difficult version, the response terms for 10 pairs were interchanged and 3 were left unchanged. We argue that the prepotent response, the original pairing, was potentiated in the audience condition. (In fact, Bond (1982) presented the pairs that were to be complex twice as often in the training phase, which should ensure that the original response is prepotent.) We argue that it is quite possible that the participants in the audience condition immediately recognized the fact that their initial, prepotent response could be wrong and were attempting to learn the correct responses. This relearning attempt led to no difference between alone and audience on the easy list (only 3 items to relearn), but to interference on the simple items embedded in the difficult list. After all, in this case, the participants had to relearn 10 of the 13 paired associates, interchanging response terms that had been correct in the original training. Mere effort would predict that had Bond (1982) gone beyond 10 trials, he would have found better performance by participants in the audience condition on both the easy and the difficult list. Of course, this account is post hoc and would have to be tested, but at least represents a plausible alternative interpretation.

(10 high frequency words and 10 low) and initial letter was a between-subjects factor.

The 20 high frequency five letter words were selected from the Kucera and Francis (1967) word frequency list (10 began with vowels and 10 began with consonants), and the 20 low frequency words were taken from Witte and Freund (2001, Experiment 3; 10 began with consonants, 10 began with vowels). We tested the frequencies of these words in a 2 (high vs. low frequency) × 2 (vowel vs. consonant) analysis of variance (ANOVA) to ensure that high frequency words were more frequent than low frequency words ($p < .0001$) and that the words beginning with vowels and consonants did not differ in frequency ($p > .90$). The presentation order of the 20 anagrams was randomized in each version of the experiment.

After completing the anagrams, the participants were asked to respond to manipulation checks for experimenter evaluation and task difficulty.

Results

Manipulation Checks

Experimenter evaluation. In Version 1, responses to the question asking to what extent the experimenter would know how well the participants performed were analyzed in a 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (frequency: high vs. low word frequency) between-subjects ANOVA. Participants in the experimenter evaluation conditions reported that the experimenter knew how well they performed to a greater extent ($M = 8.75$, $SD = 2.92$) than participants in the no experimenter evaluation conditions ($M = 4.39$, $SD = 3.31$), $F(1, 43) = 22.30$, $p < .0001$, $d = 1.44$. In Version 2, the 2 (experimenter evaluation vs. no experimenter evaluation) × 2 (vowel vs. consonant as initial letter) between-subjects ANOVA also showed that participants subject to experimenter evaluation reported that the experimenter could evaluate them to a greater extent ($M = 8.28$, $SD = 3.08$) than participants who were not subject to experimenter evaluation ($M = 3.17$, $SD = 2.44$), $F(1, 45) = 42.50$, $p < .0001$, $d = 1.94$.

Task difficulty. In Version 1, the anagram difficulty ratings were analyzed in a 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (frequency: high vs. low word frequency) ANOVA. Participants who worked on anagrams made from low frequency words rated them as more difficult ($M = 9.39$, $SD = 1.59$) than did participants exposed to anagram made from high frequency words ($M = 7.21$, $SD = 1.79$), $F(1, 43) = 18.93$, $p < .0001$, $d = 1.33$. This analysis averaged over the anagrams made from words that begin with vowels and consonants.

In Version 2, the ratings were analyzed in a 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (initial letter: vowel vs. consonant as the initial letter) between-subjects ANOVA, which showed that participants who worked on anagrams made from words that began with vowels rated the task as more difficult ($M = 9.32$, $SD = 1.73$) than did participants who worked on anagrams made from words that began with consonants ($M = 7.25$, $SD = 1.98$), $F(1, 45) = 15.03$, $p < .0001$, $d = 1.16$. This analysis averaged over the anagrams made from high and low frequency words.

Anagrams Solved

The number of anagrams solved was analyzed in a 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (word frequency: high vs. low) × 2 (initial letter: vowel vs. consonant) ANOVA. In Version 1, word frequency was a between-subjects factor and initial letter was a within-subjects factor. In Version 2, initial letter was a between-subjects factor and word frequency was a within-subjects factor. The potential for experimenter evaluation was a between-subjects factor in each version.

In Version 1, in which word frequency was a between-subjects factor, consistent with the mere effort and drive/evaluation apprehension predictions, we found a main effect for word frequency, $F(1, 43) = 24.11$, $p < .0001$, $d = 1.50$. Participants who saw anagrams formed from words of high frequency solved more of them ($M = 5.00$, $SD = 2.26$) than did participants who saw low frequency words ($M = 2.96$, $SD = 2.05$). Also consistent with the mere effort and drive/evaluation apprehension predictions, the Experimenter Evaluation × Word Frequency interaction was not reliable ($F < 1$).

This analysis also revealed an interaction between initial letter and potential for experimenter evaluation, $F(1, 43) = 9.77$, $p < .004$, $d = 0.95$. Consistent with the mere effort and drive/evaluation apprehension accounts, planned contrasts (Kirk, 1995) showed that participants subject to evaluation solved more anagrams that began with consonants ($M = 5.71$, $SD = 2.03$) than did participants who were not subject to evaluation ($M = 4.83$, $SD = 1.78$), $F(1, 43) = 4.87$, $p < .05$, $d = 0.67$, whereas participants subject to experimenter evaluation solved fewer anagrams that began with vowels ($M = 2.29$, $SD = 1.99$) than did no-experimenter-evaluation participants ($M = 3.17$, $SD = 2.15$), $F(1, 43) = 4.87$, $p < .05$, $d = 0.67$. The main effect for initial letter must be interpreted in the context of this interaction, $F(1, 43) = 80.98$, $p < .0001$, $d = 2.74$.

In Version 2, in which the manipulation of initial letter (consonant vs. vowel) was a between-subjects factor, the pattern of findings replicated the results of Version 1. Thus, once again, we found a reliable main effect for word frequency, $F(1, 45) = 22.79$, $p < .0001$, $d = 1.42$. Participants solved more anagrams formed from words of high frequency ($M = 4.49$, $SD = 2.73$) than anagrams formed from words of low frequency ($M = 2.90$, $SD = 2.41$). And, as in Version 1, the Experimenter Evaluation × Word Frequency interaction was not reliable ($F < 1$).

Once again, the analysis revealed an interaction between initial letter and potential for experimenter evaluation, $F(1, 45) = 13.31$, $p < .001$, $d = 1.09$. Consistent with the mere effort and drive/evaluation apprehension accounts, planned contrasts (Kirk, 1995) showed that participants subject to evaluation solved more anagrams that began with consonants ($M = 5.88$, $SD = 1.90$) than did participants who were not subject to evaluation ($M = 3.88$, $SD = 1.94$), $F(1, 45) = 7.52$, $p < .01$, $d = 0.82$, whereas experimenter evaluation participants solved fewer anagrams that began with vowels ($M = 1.73$, $SD = 1.89$) than did no-experimenter-evaluation participants ($M = 3.46$, $SD = 3.12$), $F(1, 45) = 5.83$, $p < .02$, $d = 0.72$. Once again, participants solved more anagrams formed from words that began with consonants ($M = 4.88$, $SD = 2.15$) than anagrams formed from words that began with vowels ($M = 2.56$, $SD = 2.67$), $F(1, 45) = 19.93$, $p < .0001$, $d = 1.33$,

but this main effect must be interpreted in the context of the Experimenter Evaluation × Initial Letter interaction.

Discussion

The manipulation check for experimenter evaluation indicated that this manipulation was successful in both versions of the experiment. The manipulation checks for task difficulty also indicated that these manipulations were successful. In Version 1, participants exposed to anagrams formed from low frequency words rated them as more difficult than did participants exposed to anagrams formed from high frequency words. In Version 2, participants exposed to anagrams made from words that began with a vowel rated them as more difficult than did participants exposed to anagrams made from words that began with consonants.

The mere effort and drive/evaluation apprehension accounts predicted a main effect for word frequency, but no interaction with experimenter evaluation, and that is exactly what we found, whether the manipulation was within subjects or between subjects. The mere effort and drive/evaluation apprehension prediction of an interaction between evaluation potential and initial letter was also supported. That is, in both Version 1 and Version 2 of the experiment, participants subject to evaluation solved significantly more anagrams that began with consonants than did participants not subject to evaluation, whereas experimenter evaluation participants solved fewer anagrams that began with vowels than did their no-experimenter-evaluation counterparts.

These findings are inconsistent with the withdrawal of effort and processing interference accounts, whether difficulty is experienced in the aggregate or at the item level. In the former case, we would have expected interactions between evaluation and whichever difficulty manipulation was in the between-subjects portion of the design (i.e., an evaluation by word frequency interaction when word frequency was the between-subjects variable and an evaluation by initial letter interaction when initial letter was the between-subjects variable). In the latter case, we expected interactions among evaluation, initial letter, and word frequency, whether the manipulation were between subjects or within subjects.

Taken together, the results from Experiment 1 provide support for the mere effort and drive/evaluation apprehension accounts over two other potential explanations, withdrawal of effort and processing interference. As noted previously, focus of attention, another potential explanation, does not make strong predictions for the anagram task. In our next set of experiments, we used the Stroop Color–Word Task (Stroop, 1935) to pit mere effort against the focus of attention account. This task also allows us to contrast the mere effort account against the explanation suggested by drive/evaluation apprehension (Cottrell, 1972; Zajonc, 1965).

Stroop Color–Word Task

The Stroop Color–Word Task (Stroop, 1935) requires participants to name the ink color of a color word. For example, they may see the word *red* printed in blue ink, and the correct response is "blue." Consistent with Baron's (1986) analysis, Huguet, Galvaing, Monteil, and Dumas (1999; see also Huguet, Dumas & Monteil, 2004) have found that social presence enhances performance on the Stroop and have argued that this facilitation was a result of the fact that social presence reduced the range of cues used by the participants. As they write, "Narrowing one's focus should indeed allow one to screen out the incorrect semantic cues and focus more exclusively on the letter color cues" (Huguet et al., 1999, p. 1013). That is, these participants see less of the word, and so, it interferes less with their response.

In their review of previous work on the Stroop, Huguet et al. (1999) cited work that they suggested shows that "arousal has been associated with *increased* [italics added] Stroop interference in past research" (p. 1012; e.g., Hochman, 1967, 1969; Pallak, Pittman, Heller, and Munson, 1975). They cited other research that shows that "distraction has been associated with *decreased* [italics added] Stroop interference in past research" (p. 1013; e.g., Houston, 1969; Houston & Jones, 1967; O'Malley & Poplawsky, 1971), as well as MacKinnon, Geiselman, and Woodward's (1985) research, which shows that coaction decreases Stroop interference. Huguet et al. (1999) commented on the contradictory nature of these findings and argued that their well-controlled experiments show that social presence reduces the amount of Stroop interference, consistent with the focus of attention interpretation.

However, we argue that the findings of the previous work are not contradictory, and, in fact, are consistent with the mere effort account. Consistent with the mere effort and drive accounts, in all cases reading the word is the prepotent (dominant) response. However, on the Stroop, unlike the RAT, the fact that this response is incorrect is quite obvious, as is the correct response. Thus, when given sufficient time, participants who are more motivated can inhibit the prepotent response and still produce the correct response more quickly than can participants who are less motivated.

In the previous research that showed increased Stroop interference, the participants had only 1 s to produce the response (Hochman, 1967, 1969; Pallak et al., 1975), and the dependent measure was the number of errors. Under these conditions, the more motivated participants did not have sufficient time to inhibit the prepotent response to read the word and make the correct response. Thus, the motivated participants made more errors than did less motivated participants. In the experiments that show decreased Stroop interference (Houston, 1969; Houston & Jones, 1967; Huguet et al., 1999; MacKinnon et al., 1985; O'Malley & Poplawsky, 1971), the dependent measure was the time required to read a whole list of color words or the time required for each individual response. Under these conditions, the motivated participants had sufficient time to inhibit the reading response and still produce the color response more quickly than did the less motivated participants.

This analysis suggests that we should be able to produce either facilitation or debilitation simply by manipulating the time available for the response. When given a limited response window (e.g., 1 s), participants subject to evaluation should make reliably more errors than should participants who are not. That is, the prepotent response will be to read the word, and given a brief response period, participants subject to evaluation will not have enough time to inhibit this response and generate the correct response. When more time is made available for the response (e.g., 2 s), few mistakes should be made and participants subject to evaluation should name the colors more quickly than should participants who are not subject to evaluation.

In contrast, the focus of attention account predicts better performance by participants subject to evaluation at each response window. Reducing the amount of time available for a response should not diminish the advantage afforded by a restricted focus of

attention. In fact, if anything, one could argue that restricted focus would lead to a greater performance advantage at the brief exposure period because these participants only see, and only need to see, part of the color word to respond correctly.

Finally, the drive/evaluation apprehension account (Cottrell, 1972; Zajonc, 1965) would predict that the presence of others simply increases drive, energizing the dominant response, reading the word. Thus, this account would predict that participants subject to evaluation would perform more poorly than would participants in the no-evaluation condition, regardless of the time available for a response.

The other two explanations, withdrawal of effort and processing interference, do not make strong predictions for performance on the Stroop. Central to each of these explanations is the participants' experience of difficulty and sense of impending failure. Overall, participants perform well on the Stroop. On the 2 s version, the error rate is extremely low (e.g., <5% in Huguet et al.'s 1999 research), but even on the 1 s version, participants correctly respond on the great majority of the trials. For example, in the condition with the poorest performance, Pallak et al.'s (1975) participants made mistakes on 16% of the trials and Hochman's (1967) participants made mistakes on 12%. Given this relatively high level of performance, there is no reason for participants to fear failure, leading to withdrawal of effort or processing interference resulting from rumination about poor performance.

Experiment 2: The 2 Second Stroop

In the first experiment, we manipulated evaluation potential and gave participants 2 s to respond to the color words and the control stimuli. When participants are provided with sufficient time to respond to the stimuli (2 s), focus of attention and mere effort accounts each predict that participants subject to evaluation should respond more quickly than should participants who are not, but for different reasons. In the focus of attention account, it is argued that the restricted focus produced by the potential for evaluation allows the participants to screen out the peripheral cue, the word, in favor of the central cue, the color. In contrast, in the mere effort account, it is argued that the potential for evaluation potentiates the prepotent (dominant) response, reading the word; but given enough time, the motivation to correct will yield a faster reaction time for the participants subject to evaluation than for those who are not. In contrast, the drive/evaluation apprehension account predicts that participants subject to evaluation should perform more poorly than should participants in the no-evaluation condition, because drive nonselectively energizes the dominant response, reading the word.

Method

Participants

Twenty-four undergraduates (50% female, 50% male) from an Introductory Psychology course at Northeastern University participated in this experiment as a means of satisfying a course requirement.

Procedure

Participants were run individually. Upon entering the lab, participants were escorted to a cubicle and seated in front of a computer. They were told that in each block of trials, they would be presented with a color word or a string of four *X*s presented in one of the four primary colors (blue, green, red, and yellow). Their task in all cases was to call out the name of the color in which the stimulus word or control string was printed, and they were to do so as quickly as possible while minimizing the number of errors. After the practice session, which consisted of four trials of color words printed in different colored ink and four trials of colored Xs presented in random order, the potential for experimenter evaluation was manipulated. Half of the participants were told that the experimenter was interested in their performance as individuals and that he or she would examine their performance at the end of the experiment but would be unable to provide individual feedback. The other half of the participants were told that the experimenter was interested in average performance and so would not look at their individual performance. Instead, at the end of the experiment, participants were asked to press the *S* key on the computer to score their performance. Then they were instructed to press the *A* key on the computer so that their score could be averaged with the performance of previous participants.

After the manipulation of experimenter evaluation, participants were told that they would perform two blocks of trials. To begin the first block, they were to wait until the experimenter left the room and to then hit the space bar. They were also told that at the end of the first block, they would be asked to press the *B* key to start the second block of trials. The experimenter then left the room and the participants performed two blocks, each consisting of 48 trials. In each block, color words were presented 24 times and a string of four *X*s was presented 24 times. Each color word was printed twice in each of the three other colors for a total of 24 trials. The 24 control trials consisted of the four *X*s printed in the four colors, six times each. The 48 trials were randomized and were followed by a second block of 48 trials, also in random order. After the first block of trials ended, participants were presented a screen, which informed them that the first block had ended and that they were to press the *B* key to start the next block of trials. The program used to run this experiment collected both the participants' verbal responses and the reaction times for these responses. Once the two trial blocks were completed, participants filled out a questionnaire, and then they were debriefed.

Results

Manipulation Check for Experimenter Evaluation

Responses to the question asking to what extent the experimenter would know how well the participants performed (1 = *not at all*; 11 = *know exactly*) were analyzed in a one-way (condition: experimenter evaluation vs. no experimenter evaluation) ANOVA. Participants in the experimenter evaluation condition believed that the experimenter knew how well they performed to a greater extent ($M = 9.00$, $SD = 2.49$) than did participants in the no experimenter evaluation conditions ($M = 3.83$, $SD = 3.41$), $F(1, 22) = 18.01$, $p < .0003$, $d = 1.81$.

Stroop Performance

The Stroop data were analyzed with 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (stimulus type:

color word vs. control stimulus) × 2 (block: block 1 vs. block 2) ANOVAs, unless otherwise noted.

Errors. Overall, the error rate was very low, <1%. However, participants subject to experimenter evaluation made fewer errors across the 48 trials (averaged over blocks) ($M = 0.10$, $SD = 0.31$) than did participants who were not subject to this evaluation ($M = 0.29$, $SD = 0.58$), $F(1, 22) = 5.10$, $p < .05$, $d = 0.96$. In addition, participants made more errors on color words ($M = 0.31$, $SD = 0.55$) than on control stimuli ($M = 0.08$, $SD = 0.35$), $F(1, 22) = 7.44$, $p < .05$, $d = 1.16$.

Latencies. Only the latencies for correct responses were submitted to analysis. However, because the overall error rate was very low, including incorrect trials does not impact the results. This analysis revealed a reliable main effect for stimulus type. Participants responded more quickly to the control stimuli ($M = 659.49$ ms, $SD = 79.92$ ms) than to the color words ($M = 777.82$ ms, $SD = 123.66$ ms), $F(1, 22) = 95.7$, $p < .0001$, $d = 4.17$, replicating the typical Stroop effect (Stroop, 1935; Huguet et al., 1999). In addition, participants responded more quickly in the first block of 48 trials ($M = 704.50$ ms, $SD = 117.19$ ms) than in the second block ($M = 732.81$ ms, $SD = 121.19$ ms), $F(1, 22) = 5.48$, $p < .03$, $d = 1.00$.

Finally, participants subject to experimenter evaluation responded more quickly ($M = 676.04$ ms, $SD = 96.45$ ms) than did participants who were not subject to this evaluation ($M = 761.27$ ms, $SD = 125.77$ ms), $F(1, 22) = 6.20$, $p < .03$, $d = 1.06$. The Experimenter Evaluation × Stimulus Type interaction approached a conventional level of significance, $F(1, 22) = 3.56$, $p = .07$, $d = 0.80$, as a result of the fact that the difference in response latency between participants who were subject and who were not subject to evaluation was greater for words ($M = 108.05$ ms) than for the control stimuli ($M = 62.41$ ms). Nonetheless this difference was reliable for both types of stimuli ($ps < .05$; Tukey honestly significant difference [HSD], Kirk, 1995). For color words, participants subject to evaluation responded in an average of 723.80 ms ($SD = 91.32$ ms), compared with the average latency of 831.85 ms ($SD = 126.69$ ms) for participants who were not subject to this evaluation, whereas for control stimuli, participants subject to evaluation responded in an average of 628.28 ms ($SD = 76.88$ ms), compared with the average latency of 690.69 ms ($SD = 71.47$ ms) for participants who were not.

Discussion

As expected, the overall error rate on the Stroop task was extremely low, <1%. Even so, participants made more errors on color words than on the control stimuli, and participants subject to evaluation made fewer errors than did those who were not.

On the primary dependent variable, latency, we found a main effect for stimulus type. Participants responded more quickly to the control stimuli than to the color word stimuli. This finding replicates the basic Stroop effect (Huguet et al., 1999; Stroop, 1935). We also found that participants subject to evaluation responded reliably more quickly than did participants who were not subject to this evaluation. There was a marginal Experimenter Evaluation × Stimulus type interaction produced by the fact that although participants subject to evaluation were faster on both the control and the color words, the difference was greater for color words. It should be noted that these findings are not the result of a speed–accuracy trade off. That is, participants subject to evaluation were not faster because they sacrificed accuracy for speed. They made fewer errors and were faster in making their responses than their no-evaluation counterparts.

Finding that participants subject to evaluation responded more quickly than did no-evaluation participants is consistent with both the focus of attention and mere effort accounts. The focus of attention explanation suggests that these findings result from the fact that participants subject to experimenter evaluation have a restricted focus of attention, which makes the task easier for them. In contrast, in the mere effort account, it is argued that the prepotent response is to read the word and that the potential for experimenter evaluation potentiates this response. However, the heightened motivation of these participants also prompts them to attempt to inhibit this incorrect response and to make the correct response, and the 2 s response window provides them with enough time to do so.

These findings are not consistent with the drive/evaluation apprehension account. Energization of the dominant response alone would predict poorer performance for participants subject to evaluation than for participants in the no-evaluation condition, and this was not the case. However, in the mere effort account, as in drive/evaluation apprehension, it is argued that the prepotent response is potentiated, and this potentiation would be revealed if the response time available were not sufficient for correction. This prediction was tested in the next experiment.

Experiment 3: Stroop Task With Limited Response Time

In this experiment, the procedures were the same as those that were used in Experiment 2. The only difference was the time provided for the response. In one condition, participants were given 1 s to respond, as in previous Stroop research that Huguet et al. (1999) cites as showing increased Stroop interference (e.g., Hochman, 1967, 1969; Pallak et al., 1975). In a second condition, participants were given 750 ms to make their responses to test the effect of increased task demand on the participants' performance.

The mere effort hypothesis predicts that when time to respond is significantly decreased (e.g., from 2 s to 1 s), participants subject to experimenter evaluation will make more errors than will participants who are not. That is, experimenter evaluation participants will not have sufficient time to inhibit the prepotent response and produce the correct response when given only 1 s. Decreasing the time available for the response even more (e.g., from 1 s to 750 ms) should exacerbate this tendency. The drive/evaluation apprehension account would also predict energization of the dominant response, leading to poorer performance by participants subject to evaluation than by those who are not, at both 1 s and 750 ms. In contrast, the focus of attention account predicts that participants subject to evaluation would still outperform those that are not. There is no reason that a brief response period should interfere with the advantage provided by the restricted focus of attention. In fact, restricted focus of attention could represent an even greater advantage at a briefer exposure period.

Method

Participants

Forty-nine (51% female, 49% male) Northeastern University undergraduates participated in this experiment as a means of satisfying a course requirement.

Procedure

By the same procedures as in the previous experiment, participants were randomly assigned to either the experimenter evaluation condition or the no experimenter evaluation condition. Crossed with this manipulation, participants were randomly assigned to either a 1 s or a 750 ms condition. The dependent measure was the number of errors made.

Results

Manipulation Check for Experimenter Evaluation

Responses to the question asking to what extent the experimenter would know how well the participants performed were analyzed in a 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (response window: 1 s vs. 750 ms) ANOVA. Participants in the experimenter evaluation condition reported that the experimenter knew how well they performed to a greater extent ($M = 8.79$, $SD = 2.73$) than did participants in the no experimenter evaluation conditions ($M = 4.67$, $SD = 3.92$), $F(1, 45) = 17.42$, $p < .0001$, $d = 1.24$.

Stroop Performance

The Stroop data were analyzed with 2 (condition: experimenter evaluation vs. no experimenter evaluation) × 2 (stimulus type: color word vs. control stimulus) × 2 (response window: 750 ms vs. 1 s) × 2 (block: Block 1 vs. Block 2) ANOVAs, unless otherwise noted.

Errors. Analysis of errors revealed a main effect for stimulus type. Participants made more errors on color words ($M = 4.42$, $SD = 3.80$) than on control stimuli ($M = 0.72$, $SD = 1.14$), $F(1, 45) = 95.89$, $p < .0001$, $d = 2.92$. This analysis also revealed a main effect for response window, $F(1, 45) = 29.20$, $p < .0001$, $d = 1.61$. Participants in the 750 ms condition made more errors ($M = 3.81$, $SD = 3.83$) than did participants in the 1 s condition ($M = 1.38$, $SD = 2.26$). However, these main effects must be interpreted in the context of the reliable Stimulus Type × Response Window interaction, $F(1, 45) = 15.71$, $p < .0003$, $d = 1.18$. This interaction was produced by the fact that participants in the 750 ms condition did not make reliably more errors when responding to the control stimuli ($M = 1.21$, $SD = 1.34$) than did participants who were given 1 s to respond ($M = 0.26$, $SD = 0.63$; $p > .25$). However, participants in the 750 ms condition did make reliably more errors when responding to color words ($M = 6.42$, $SD = 3.75$) than did participants in the 1 s condition ($M = 2.50$, $SD = 2.72$), $p < .05$ (Tukey HSD, Kirk, 1995).

As in the previous Stroop research in which response time was limited (e.g., Hochman, 1967, 1969; Pallak et al., 1975), participants subject to experimenter evaluation made more errors ($M = 3.07$, $SD = 3.78$) than did participants in the no experimenter evaluation condition ($M = 2.05$, $SD = 2.77$), $F(1, 45) = 5.56$, $p < .03$, $d = 0.70$. However, this main effect must be interpreted in light of the reliable Stimulus Type × Experimenter Evaluation interaction, $F(1, 45) = 5.59$, $p < .03$, $d = 0.70$. Participants in the experimenter evaluation condition made more errors on color words ($M = 5.34$, $SD = 4.14$) than did participants in the no experimenter evaluation condition ($M = 3.46$, $SD = 3.17$), $p < .05$ (Tukey HSD, Kirk, 1995), whereas on control stimuli, participants subject to experimenter evaluation made no more errors ($M = 0.80$, $SD = 1.09$) than did participants who were not subject to this evaluation ($M = 0.65$, $SD = 1.19$, $p > .20$). The three-way interaction, Stimulus Type × Experimenter Evaluation × Response Window, was not significant ($p > .80$).

Latencies. As a result of the differences in accuracy, the condition means are based on different numbers of trials, making any comparisons of reaction time suspect. In addition, on inhibition tasks like the Stroop, when response time is limited, differences are typically reflected in accuracy (e.g., Hochman, 1967; Pallak et al., 1975) but are reflected in speed when the time provided for a response is essentially unlimited (Exp. 3; see also Huguet et al., 1999; MacKinnon et al., 1985; O'Malley & Poplawsky, 1971).

In any event, the analysis of the latencies for correct responses revealed no reliable effects for evaluation. There was a reliable main effect for response window, $F(1, 45) = 10.13$, $p < .01$, $d = 0.95$, and a main effect for stimulus type, $F(1, 45) = 173.72$, $p < .0001$, $d = 3.93$. However, these main effects must be interpreted in the context of the Stimulus Type × Response Window interaction, $F(1, 45) = 25.08$, $p < .0001$, $d = 1.49$. A Tukey HSD (Kirk, 1995) shows that although the differences between color words and control stimuli were significant at both 1 s ($M_{\text{color words}} = 749.32$, $SD = 65.40$; $M_{\text{control}} = 652.67$, $SD = 58.64$, $p < .05$), and 750 ms ($M_{\text{color words}} = 676.14$, $SD = 36.33$; $M_{\text{control}} = 632.51$, $SD = 58.82$, $p < .05$), the difference between these groups was greater at 1 s than at 750 ms.

There was also a main effect for blocks, $F(1, 45) = 6.71$, $p < .02$, $d = 0.77$, which must be interpreted in the context of the Block × Response Window interaction, $F(1, 45) = 15.91$, $p < .0001$, $d = 1.19$. A Tukey HSD (Kirk, 1995) showed that when the response window was 1 s, response times were faster in Block 1 ($M = 690.29$, $SD = 82.11$) than in Block 2 ($M = 712.31$, $SD = 74.51$), $p < .05$, whereas at 750 ms, there was no difference ($M_{\text{Block 1}} = 656.68$, $SD = 52.48$; $M_{\text{Block 2}} = 651.97$, $SD = 54.65$, $p > .20$).

Discussion

On the primary dependent variable, errors, we found a main effect for stimulus type. Participants made more errors when presented with a color word than when presented with a control stimulus, thus replicating the Stroop effect. We also found that participants in the 750 ms condition made more errors than did participants in the 1 s condition. Of course, these main effects must be considered in light of the Stimulus Type × Response Window interaction produced by the fact that participants given 750 ms to respond made more errors when responding to the color words than did their 1 s counterparts, but response window did not impact errors on the control stimuli.

There was also a main effect for evaluation potential. Participants subject to evaluation made more errors than did those not subject to this evaluation. However, this effect must be interpreted in the context of the reliable Stimulus Type × Experimenter Evaluation interaction. Participants in the evaluation condition made more errors on color words than did participants in the no-evaluation condition, whereas on the control stimuli there was no difference as a function of evaluation potential.

Finding that participants subject to evaluation make more errors on color–words than do no-evaluation participants is consistent with mere effort and drive/evaluation apprehension predictions. Each of these accounts would contend that reading color words is the prepotent (dominant), but incorrect, response and that the motivation produced by the potential for evaluation potentiates this response. However, the drive/evaluation apprehension account only predicts this energization, whereas the mere effort account also predicts that participants subject to evaluation will be motivated to produce the correct answer. At the brief response windows (Experiment 3), these participants do not have sufficient time to inhibit the prepotent response and produce the correct response, and as a result, they perform more poorly than do no experimenter evaluation participants. However, at the longer response window (Experiment 2), they are able to inhibit the incorrect response and produce the correct response faster than do no experimenter evaluation participants.

Finding that participants subject to evaluation respond more quickly than do participants in the no-evaluation condition in Experiment 2 is consistent with the focus of attention prediction. However, the findings of Experiment 3 are not. If anything, a briefer display period should represent an advantage for participants with a narrowed focus of attention. Instead we find that participants subject to evaluation make more errors than do participants who are not. Thus, in this test of the focus of attention and the mere effort explanations, we find support for the mere effort account.

Recently, Muller and Butera (2007; see also Muller, Atzeni, & Butera, 2004) have extended Huguet et al.'s (1999) focus of attention account by using a perceptual task, developed by Treisman and her colleagues (Treisman & Paterson, 1984; Treisman & Schmidt, 1982), which, unlike the Stroop, does not involve verbal processes. In that research, self-evaluation threat is created by leading participants to believe that they are performing more poorly than a coactor (upward social comparison) or that they are performing with a coactor about whose performance they will be given no information (mere coaction in the study's terminology). Under these circumstances, participants ruminate about the existing (upward comparison) discrepancy or the potential (mere coactor) discrepancy between their performance and their standards. These ruminations consume attentional resources that would otherwise be devoted to processing peripheral cues, resulting in attentional focusing on the central cues. On the other hand, when participants are not faced with a coactor (without social comparison) or find that they are outperforming the coactor (downward social comparison), there is no self-evaluation threat and, thus, no attentional focusing.

To test this account, Muller and his colleagues (Muller, Atzeni, & Butera, F., Experiments 1 & 2; Muller & Butera, 2007, Experiments 1–4) used an illusory conjunction task (e.g., Triesman, 1988) and, consistent with their argument, found that participants subject to self-evaluation threat (upward social comparison and mere coaction) reported fewer illusory conjunctions than did participants not subject to self-evaluation threat (downward social comparison). However, as Muller and Butera (2007) note, "the lower conjunctive error rate may not be a specific effect of attentional focus but a generic social facilitation effect due to an increase in effort" (p. 205). In fact, the pattern of results produced in these experiments is not inconsistent with the mere effort account. That is, participants subject to the potential for self-evaluation threat would be more motivated to perform well and would thus exert greater effort on the task.

To rule out this type of motivational explanation for their findings, Muller and Butera (2007) conducted a fifth experiment, the findings of which they argued support the attentional focus account but do not support the motivational explanations. Participants were asked to look at a fixation point that was displayed for 1,000 ms. A dot then flashed approximately 7° from the screen's center in one of four locations for 30 ms, followed by a display of four letters (three *Q*s and one *O*) that formed a square. The participants' task was to identify the location of the *O*. On half of the trials, the dot flashed in the location where the *O* would appear. On the other half of the trials, the dot flashed in a location where a *Q* would appear. That is, half of the time the dot was a valid cue as to the location of the *O*, whereas the other half of the time the dot was an invalid cue. According to Muller and Butera (2007), the central cue is the array of letters and the peripheral cue is the dot. Thus, the attentional focus account predicted that participants experiencing the self-evaluation threat produced by upward social comparison would focus on the central cue and be less affected by the peripheral cue than would participants not experiencing threat (downward social comparison).

Consistent with their argument, Muller and Butera (2007) found that the reaction times of participants in the upward social comparison condition did not differ as a function of cue validity, whereas the reaction times of participants in the downward comparison condition did. Nonetheless, we argue that the overall pattern of results is not consistent with Muller and Butera's account. In the downward social comparison condition, valid cues led to faster reaction times (514 ms) than did invalid cues (534 ms). If the participants in the upward social comparison condition ignored the peripheral cues, their reaction times should have fallen within the bounds set by these conditions. That is, if they are ignoring the valid cue, their reaction times should be slower than the reaction times of the participants in the downward social comparison condition who are attending to the cue, and if they are ignoring the invalid cue, their reaction times should be faster than the reaction times of these participants.[2] However, this was not the case. In the upward social comparison condition, on the trials on which the cue was valid, the mean reaction time was 567 ms, and when invalid, 561 ms. Finding reaction times outside the bounds set by the performance of the participants in the downward comparison condition is not consistent with Muller and Butera's focus of attention account.

[2] Muller and Butera (2007) argued that in the downward social comparison condition "invalid cues should only lower reaction times slightly because without any cue, the serial search would only have one chance out of four to start in the right location against zero chances out of four when attention is attracted by the invalid cue" (p. 205). As a result, the improvement in reaction time produced by attending to the valid cue should be greater than the reduction in reaction time produced by attending to the invalid cue. Nonetheless, if participants in the downward social comparison condition respond more slowly than did upward social comparison participants on 25% of the trials and on the remaining 75% of the trials, the two groups are equal; the overall the reaction times of the downward social comparison participants should be slower than the reaction times for upward social comparison participants.

On the other hand, these findings are not inconsistent with the mere effort account. Mere effort predicts that self-evaluation threat should motivate participants to perform well, which should potentiate the prepotent response. Previous research has shown that abrupt visual onsets, like the onset of the cue in Muller and Butera's (2007) task, attract participants' attention (e.g., Remington, Johnston, & Yantis, 1992; Yantis & Jonides, 1984; Yantis & Jondies, 1990), and self-evaluation threat should potentiate this prepotent tendency. Thus, the mere effort account would predict that participants in the upward social comparison condition would be more likely (not less likely) to look toward the cue than the control participants. Of course, on any given trial, looking at the dot could lead participants to the correct answer (the *O*), or to an incorrect answer (a *Q*).

On this task, as on the Stroop, whether the answer is correct or not is quite obvious, and the mere effort account would predict that participants subject to self-evaluation threat should be highly motivated to produce the correct answer. However, the fact that on any given trial the cue may be a valid or an invalid indicator of the location of the correct answer may have led these participants to be cautious in their responses, to take the time necessary to make certain that they had the correct answer. This tendency toward caution could be accentuated by the fact that participants in the upward social comparison condition were told that they scored 65 out of a 100 in their initial performance of the task, whereas the coactor scored 80. They were told nothing about the extent to which speed versus accuracy contributed to this score. Because accuracy scores are given as a value out of 100, upward social comparison participants could well have concluded that they performed more poorly than the coactor because their responses were inaccurate, which would also lead them to be cautious in their responses.

The effect of this caution could be to slow down these participants' responses and to eliminate the difference in reaction time between trials with valid cues and those with invalid cues. If this is the case, if we remove the ambiguity in the meaning of the cue and give the participants no reason to believe that they are inaccurate, we should be able to see both the effect of the potentiation of the prepotent response and the motivation to correct. The antisaccade task (Hallet, 1978) serves this purpose well.

Antisaccade Task

On the antisaccade task, a participant is asked to fixate a cross that appears in the center of the visual display and to respond to a target presented randomly on one side of the display or the other. However, before the target appears, a cue (a white square) is presented on the opposite side of the display. Participants are instructed to not look at this cue but rather to look to the opposite side of the display where the target will appear. However, there is a reflexivelike, prepotent tendency to look at the cue that must be inhibited to optimize performance (see Figure 1). Thus, this task shares many of the features of the task used by Muller and Butera (2007). For instance, both tasks begin with the presentation of a central fixation, followed by an abrupt onset peripheral cue, and then the target (central cue). Also, attention to the peripheral cue is not necessary for target identification in either task, and each task requires that participants shift their visual attention to the target's location to respond accurately. However, unlike Muller and Butera's cue, on the antisaccade task, the peripheral cue is always on the side opposite to the one on which the target will appear.

Jamieson and Harkins (2007) used this task to test a mere effort account of the effect of stereotype threat on performance. Stereotype threat, like the potential for evaluation, arouses participants'

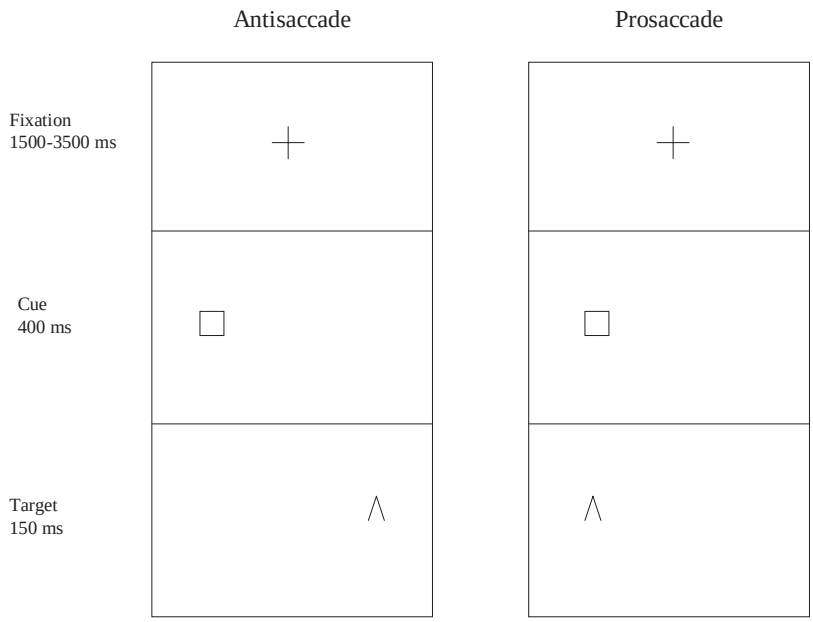

Figure 1. Sequence of events for the antisaccade and prosaccade tasks. Each frame represents what was displayed on the monitor for the period of time shown to the left of the figure. The target appeared in one of three orientations: pointing up (shown), to the right, or to the left.

concern about their ability to perform well on a task. Thus, in the mere effort account, it is argued that stereotype threat should produce the same basic pattern of findings on the antisaccade task as is produced by the potential for evaluation on other inhibition tasks like the Stroop. When not given sufficient time to correct for the prepotent tendency (i.e., at a brief display time), the more motivated threat participants should be less accurate than should controls in their ability to correctly identify target orientation. However, when the display time is increased enough to allow enough time for correction, stereotype threat participants should be able to respond to the target more quickly than should controls, as a result of increased motivation to perform well, and this is what Jamieson and Harkins (2007) found.

Jamieson and Harkins (2007) used an eye tracker to conduct a more fine-grained analysis of performance on this antisaccade task at a display time that permitted correction. Under these conditions, the mere effort account predicted that the participants under threat would look the wrong direction, toward the cue, more often than would participants in the control group, because the motivation to perform well potentiates the prepotent response. At this point, if the participants have failed to inhibit the reflexive saccade, their eyes are at the cue and they must launch a corrective saccade to get to the target site. If they have successfully inhibited the saccade, they must launch a correct saccade to the target site from the fixation point. Because correct and corrective saccades are each an "extreme example of a voluntary saccade" (Sereno, 1992, p. 92), the motivation to correct should reduce the latency to launch each type of saccade, and, as a result, the evaluation participants should launch these saccades faster than should control participants. Finally, after the participants' eyes arrive at the target area, the participant must determine the target's orientation and press the appropriate response key. When the participants see the target, the mere effort account predicted that the greater motivation of participants subject to stereotype threat would lead them to respond more quickly than would participants in the control condition.[3] Jamieson and Harkins (2007) found support for each of these predictions.

Experiment 4: Antisaccade Task With Eye Tracking

In Experiment 4, we used the antisaccade task to test the mere effort account against Muller and Butera's (2007) focus of attention account. Muller and Butera (2007) argued that the cue (the dot) did not affect the performance of participants subject to self-evaluation threat because they focused on the central cue and were able to inhibit any tendency to attend to this peripheral cue. This leads to the prediction that participants subject to evaluation should focus their attention on the central cue, the target, and should launch fewer saccades toward the peripheral cue (the box) than should participants in the no-evaluation condition. Looking the wrong way less should result in better performance by participants subject to evaluation than by participants in the no-evaluation condition.

In contrast, the mere effort account predicts that the potential for evaluation will potentiate the prepotent response (looking at the box), leading to more, not fewer, reflexive saccades than are produced by participants in the no-evaluation condition. However, in mere effort, it is also argued that participants subject to evaluation are motivated to report the orientation of the arrow as quickly as possible. As a result, they should launch correct (after successful inhibition) and corrective (after they have looked the wrong way) saccades faster than should participants in the no-evaluation condition (see Figure 2). And when the participants subject to evaluation see the target, as a result of their motivation to perform well, they should press the response key more quickly than should no-evaluation participants. As a result, even though the evaluation participants look the wrong way more often than do no-evaluation participants, they will end up outperforming these participants.

It should be noted that the focus of attention account predicts none of these specific effects. Certainly it would not predict that participants subject to evaluation would look the wrong way (toward the peripheral cue) more than would no-evaluation participants. However, this account would also not predict faster launch times for correct and corrective saccades or faster adjusted reaction times. These volitional behaviors are produced by the evaluation participants' motivation to perform well. The focus of attention account predicts that evaluation participants will outperform participants in the no-evaluation condition simply by virtue of the fact that they look at the peripheral cue less often than do the latter participants.

Method

Participants

Sixty Northeastern University students participated in this experiment in exchange for class credit. All participants reported normal vision or corrected-to-normal vision, but none wore eyeglasses because they would interfere with the accuracy of the oculometer.

Tasks and Apparatus

Each participant was seated in front of a 17 in. (43.18 cm) monitor in a small room. Stimulus presentation, key press timing, and the accuracy of the responses were controlled by a computer. Participants' heads were stabilized throughout the experiment by a chin rest positioned 54 cm from the monitor.

Participants completed two eye-movement tasks, the antisaccade and the prosaccade tasks. On the antisaccade task (see Figure 1), each trial began with the presentation of a fixation cross, subtending 1° of visual angle, in the center of the screen for a

[3] These adjusted reaction times differ from the usual reaction times in that the latter represent the time taken to respond measured from when the target appears (400 ms after cue presentation). However, on some trials, the eyes arrive in the target area after the target has already appeared, whereas on other trials, the eyes arrive before the target appears. In the latter cases, no adjustment is necessary. Reaction time measured from the 400 ms mark makes sense, but in the former, the reaction time includes time when the participants could not have responded because their eyes have not yet arrived at the target area. Therefore, to isolate the key press component, for each participant on each trial in which the saccade reached the target after its presentation, the amount of time by which it came after target presentation was subtracted from the total reaction time for that trial. On those trials on which the eyes arrived at the target area before the target appeared, the reaction time was measured from the 400 ms mark. The average adjusted reaction time for each participant was then computed and entered in the analysis.

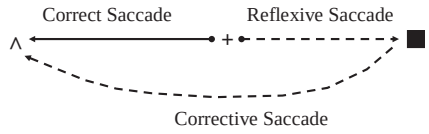

Figure 2. Response maps for different types of saccadic eye movements on antisaccade trials. The solid line represents trials on which participants first make a correct saccade toward the target. The broken line represents trials on which participants first make a reflexive saccade and then generate a corrective saccade back toward the target.

randomly determined interval ranging from 1,500 ms to 3,500 ms. The cue, a white square that subtended 0.5° of visual angle, was then presented 11° either to the left or the right of the fixation cross for 400 ms. When the cue was extinguished, the target, an arrow also subtending 0.5° of visual angle, then appeared on the opposite side of the screen from the cue, 11° from the center fixation cross. The target was presented in one of three orientations: pointing up, to the left, or to the right. The target was displayed for 150 ms, after which a mask, another white square subtending 0.5° of visual angle, appeared in its place.[4] The mask remained until the participant responded with a key press. If no response was made, the mask was removed after 1,500 ms, and the next trial began after a 1,750 ms intertrial interval.

Participants were instructed to look at the fixation cross in the center of the screen and to respond to a target presented randomly on one side of the display or the other. However, before the target appeared, a cue (a white square) would be presented on the opposite side of the display. They were instructed to not look at this cue but rather to look to the opposite of the display where the target would appear. They were to indicate the orientation of the target as quickly and accurately as possible by pressing the corresponding arrow key on a keyboard (left, up, or right). Cue side (left or right) and arrow direction were randomized across trials.

As shown in Figure 1, the prosaccade task was identical to the antisaccade task, except that the target (the arrow) was presented on the same side of the screen as the cue (the white square). Participants were instructed to look toward the cue and identify the orientation of the target that appeared in its place. The prepotent tendency to look toward the peripherally flashed cue is correct on prosaccade trials, whereas on the antisaccade task, this prepotent response is incorrect. Thus, prosaccade trials are structurally similar to antisaccade trials but do not require the inhibition and/or correction of prepotent responses.

Participants completed six practice trials prior to the beginning of each task and then completed 74 antisaccade or prosaccade trials. Task order was counterbalanced across participants. As is common in antisaccade research (e.g., Roberts, Hager, & Heron, 1994; Stuyven, Van der Goten, Vandieraendonck, Claeys, & Crevits, 2000), participants did not receive feedback after each trial.

Eye-movement data were collected with a Dr. Bouis infrared oculometer (Dr. Bouis Devices, Karlsruhe, Germany) interfaced with the computer, while head position was stabilized with a chin rest. The oculometer measured eye position by projecting an infrared light into the eye at an intensity limited to 3×10^{-4} W/cm^2 and calculating the angular disparity between pupil reflectance and maximum corneal reflectance. The resolution was only limited by the fact that the infrared light illuminating the eye was pulsed at 4 kHz. Thus, the oculometer permitted eye position to be tracked with a resolution of 0.1°, which is ideal for measuring small eye movements such as saccades (Bach, Bouis, & Fischer, 1983). To ensure that the oculometer remained calibrated for luminance and spatial accuracy throughout the experiment, an onscreen calibration test was presented every 20 trials.

This calibration test was conducted by an experimenter seated in the room with the participant but out of view of the screen. When the calibration screen appeared, participants were instructed to inform the experimenter, who then conducted the calibration test. Because participants were required to inform the experimenter when the calibration test appeared, it was obvious that the experimenter could not view the computer screen and evaluate performance.

Procedure

Participants were brought into the lab one at a time. After consent was obtained, participants were given an overview of the eye-tracking equipment and verbal instructions for the saccade tasks, followed by six practice trials. Upon completion of the practice trials, the experimenter implemented the evaluation manipulation. In the experimenter evaluation condition, participants were told that the experimenter was interested in their performance as an individual and that after each task the experimenter would be evaluating their performance while they completed a questionnaire in the adjoining room. In the no experimenter evaluation condition, participants were told that the experimenter was interested in people's performance in general and that the computer would average their scores automatically with the scores of all of the previous participants upon the completion of each task. In the no experimenter evaluation condition, participants also completed questionnaires in the adjoining room. Both groups were instructed to perform each trial as quickly and accurately as possible. Participants completed questionnaires after each saccade task.

Data Preparation

Filters were used prior to data analysis to ensure that eye movements recorded by the eye-tracker represented responses to the stimuli and were not random movements. Prior to beginning each trial, participants were required to fixate on a center fixation cross. If in the period of 200 ms preceding the onset of the cue, a

[4] To test the predictions for the mere effort account, the target must be displayed long enough for the motivation to correct to have an opportunity to play a role. That is, if the display period is too short, participants subject to evaluation will not be able to recover quickly enough to see the target. Therefore, we conducted a pilot study without eye tracking to determine the appropriate display period for the target. We first set the display time at 150 ms because this exposure time has been used in previous antisaccade research (e.g., Roberts et al., 1994), and we found that at this value, participants subject to evaluation had faster reaction times than did no-evaluation participants with no sacrifice in accuracy. As a result, we used this display time in Experiment 4. Of course, this outcome is consistent with the predictions of both mere effort and focus of attention accounts. It should also be noted that Jamieson and Harkins (2007) found that control participants outperformed stereotype threat participants at 150 ms (accuracy) and that a display time of 250 ms was required for stereotype threat participants to outperform control participants (equal accuracy and faster reaction times). We return to this issue in the Discussion section.

participant's eye position did not vary by more than 2.82° (50 pixels), then that trial was considered as having a valid baseline. If gaze strayed more than 2.82° from the center of the central position during this 200 ms pretrial window, then that trial was considered as having a bad baseline and was excluded from the analysis. A total of 3.78% of the total number of trials across the prosaccade and antisaccade tasks were excluded due to bad baselines.

Trials on which participants initiated saccades 80 ms or less after the presentation of the cue were considered anticipatory (e.g., Crevits & Vandieraendonck, 2005; Ford, Goltz, Brown, & Everling, 2005) and were excluded from the analyses. Additionally, saccades beginning at 1,000 ms or more after the presentation of the cue were excluded from the data analyses because these eye movements could not have been initiated in response to either the cue or the target because both stimuli had been previously extinguished. Use of these criteria resulted in the exclusion of another 6.22% of the trials. Thus, a total of 10% of trials were excluded from the analyses as a result of poor baselines and threshold and limit violations. The percentage of excluded trials did not differ by condition ($ps > .20$). In addition, previous antisaccade research with eye tracking measures has excluded approximately the same percentage of trials (e.g., Kane, Bleckley, Conway, & Engle, 2001; Unsworth, Schrock, & Engle, 2004).

Eye movements were classified as saccades if participants shifted their gaze position by more than 4.25°, however movements less than 4.25° were uncommon as participants exhibited a tendency to generate consistent saccadic movements ($M = 10.89°$, $SD = 3.06°$) to either the target or the cue, which were each located 11° from the center of the computer screen. Participants' average saccade velocity for an 11° eye movement ($M = 221°/s$) fell below peak human saccade velocity for 11° eye movements and was within the normal range for eye movements of this magnitude (e.g., Montagnini & Chelazzi, 2005).

Results

Unless otherwise specified, data were analyzed in 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) × 2 (task: antisaccade vs. prosaccade) ANOVAs. Evaluation and task order were analyzed as between-subjects effects, whereas task was analyzed as a within-subjects effect.

Manipulation Checks for Evaluation

Participants subject to the experimenter evaluation manipulation reported that the experimenter could evaluate their performance to a greater extent ($M = 10.32$, $SD = 7.53$) than did participants in the no-evaluation condition ($M = 3.20$, $SD = 2.90$), $F(1, 56) = 46.77$, $p < .001$, $d = 1.83$.

Performance

Accuracy. Participants were more accurate on prosaccade trials ($M = 99.17\%$, $SD = 1.26\%$) than on antisaccade trials ($M = 96.03\%$, $SD = 4.83\%$), $F(1, 56) = 22.72$, $p < .001$, $d = 1.27$. This effect is common in research with the prosaccade and the antisaccade tasks (e.g., Jamieson & Harkins, 2007; Roberts et al., 1994; Unsworth et al., 2004) and is expected because on the prosaccade task, unlike the antisaccade task, good performance does not require the inhibition of the prepotent response tendency.

Terminal reaction time. Replicating previous research (e.g., Roberts et al., 1994), participants responded to target orientation more quickly on prosaccade trials ($M = 411.24$ ms, $SD = 68.18$ ms) than on antisaccade trials ($M = 457.94$ ms, $SD = 135.65$ ms), $F(1, 56) = 21.44$, $p < .001$, $d = 1.24$. As previously noted, the antisaccade task requires the inhibition of a prepotent response, whereas the prosaccade task does not. Thus, slower reaction times are expected on the antisaccade task.

Participants subject to evaluation responded to target orientation more quickly ($M = 378.64$ ms, $SD = 54.86$ ms) than did those not subject to evaluation ($M = 492.48$ ms, $SD = 107.32$ ms), $F(1, 56) = 28.37$, $p < .001$, $d = 1.42$. However, this main effect must be interpreted in the context of a Task × Condition interaction, $F(1, 56) = 13.10$, $p = .001$, $d = 0.97$. A Tukey HSD (Kirk, 1995) test shows that although participants subject to evaluation responded more quickly than did their no-evaluation counterparts on both antisaccade ($M_{\text{Evaluation}} = 383.79$ ms, $SD_{\text{Evaluation}} = 67.08$ ms; $M_{\text{No evaluation}} = 534.65$ ms, $SD_{\text{No evaluation}} = 146.54$ ms; $p < .05$) and prosaccade trials ($M_{\text{Evaluation}} = 373.48$ ms, $SD_{\text{Evaluation}} = 42.63$ ms; $M_{\text{No evaluation}} = 450.31$ ms, $SD_{\text{No evaluation}} = 68.09$ ms; $p < .05$), the difference between these groups was greater on antisaccade trials than on prosaccade trials ($p < .05$).

Eye-Movement Measures

Analyses were conducted on the three types of saccades produced on the antisaccade task: reflexive saccades, corrective saccades, and correct saccades (see Figure 2), and on the eye movements produced on the prosaccade task. Adjusted reaction time data (reaction times adjusted for time of arrival at the target area) for antisaccades and prosaccades were also analyzed.

Reflexive saccades. The percentage and latency of reflexive saccades were analyzed in 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) ANOVAs, with condition and task order analyzed as between-subjects factors. This analysis included all the antisaccade trials that met the inclusion criteria, whether the trial ended with a correct response or not. The mere effort account predicts that participants subject to evaluation will launch a greater number of incorrect reflexive saccades than will no-evaluation participants. These predictions hold whether the trial culminates in a correct response or not.

Consistent with mere effort predictions, participants in the evaluation condition launched reflexive saccades on a greater percentage of the trials ($M = 46.90\%$, $SD = 25.81\%$) than did no-evaluation participants ($M = 29.33\%$, $SD = 23.38\%$), $F(1, 56) = 7.50$, $p = .008$, $d = 0.73$. There was also a trend for evaluation participants to launch these saccades more quickly ($M = 136.87$ ms, $SD = 30.91$ ms) than did participants in the no-evaluation condition ($M = 174.22$ ms, $SD = 129.46$ ms), $F(1, 56) = 2.14$, $p = .15$, $d = 0.39$.

Corrective saccades. The latencies of corrective saccades were analyzed in a 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) ANOVA, with condition and task order as between-subjects effects. This analysis included the antisaccade trials that met the inclusion criteria and that were correctly answered because we are attempting to account

for differences in reaction time on trials for which the response was correct. In fact, however, as noted above, the participants correctly identified the orientation of the target on 96.03% of the antisaccade trials, and evaluation did not affect their ability to do so ($p > .25$). As a result, including the few incorrect trials makes no difference in the pattern of the results.

This analysis showed that participants subject to evaluation launched corrective saccades more quickly ($M = 356.03$ ms, $SD = 59.02$ ms) than did no-evaluation participants ($M = 405.64$ ms, $SD = 81.84$ ms), $F(1, 53) = 6.87$, $p = .011$, $d = 0.72$ (see Figure 3).

Correct saccades. The latencies for correct saccades were analyzed in a 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) ANOVA, with condition and task order as between-subjects effects. Once again, we excluded the few trials on which the participants answered incorrectly. Participants subject to evaluation launched correct saccades more quickly ($M = 283.19$ ms, $SD = 70.42$ ms) than did no-evaluation participants ($M = 333.36$ ms, $SD = 73.54$ ms), $F(1, 55) = 10.86$, $p = .01$, $d = 0.89$ (see Figure 3).

Prosaccades. On the prosaccade task, the cue and target appear on the same side. Prosaccades are eye movements launched in the direction of the cue and target for this task. The latencies of these saccades were analyzed in a 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) ANOVA, with condition and task order as between-subjects effects. Once again, we used only trials that met the inclusion criteria and that were answered correctly. However, participants answered correctly on 99.17% of the prosaccade trials, so excluding the few incorrect trials did not impact the results. This analysis did not reveal any reliable effects (ps $> .20$).

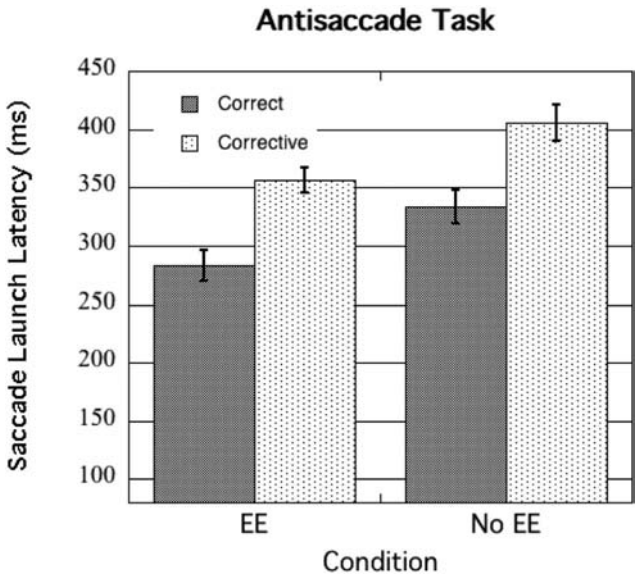

Figure 3. Saccade launch latencies for correct and corrective saccades as a function of experimenter evaluation condition in Experiment 4. The error bars are standard errors of the mean for the respective conditions. EE = Experimenter Evaluation.

Adjusted Reaction Times

Antisaccade task. To test the hypothesis that the motivation to press the response key could contribute to performance on the antisaccade task, we had to take into account the time at which the participants' eyes arrived at the target area. If their eyes arrived before the target even appeared (400 ms from the beginning of the trial) then no adjustment was necessary. The participant was looking at the target area when the target appeared (at the 400 ms mark), and reaction time as measured from the 400 ms mark until the key press was appropriate. However, if, for example, the eyes did not reach the target until the 450 ms mark, reaction time measured from the 400 ms mark would include 50 ms in which the participant could not have responded. Therefore, to isolate the key press component, in those cases in which the participant's eyes arrived at the target area prior to the target's appearance, we left the reaction time unchanged (i.e., measured from the 400 ms mark in the trial). In those cases in which the saccade reached the target after its presentation, we subtracted from the reaction time the amount of time by which it came after. This procedure was followed for each trial for each person, and the resulting adjusted reaction time scores were averaged for each person.

The adjusted reaction times were analyzed in a 2 (condition: evaluation vs. no-evaluation) × 2 (task order: antisaccade first vs. prosaccade first) × 2 (type of saccade: corrective vs. correct) ANOVA, with condition and task order as between-subjects factors and type of saccade as a within-subjects factor. The adjusted reaction times for participants subject to evaluation were significantly faster ($M = 359.79$ ms, $SD = 119.38$ ms) than were those for participants in the no-evaluation condition ($M = 485.48$ ms, $SD = 152.86$ ms), $F(1, 52) = 17.92$, $p < .001$, $d = 1.17$. This finding is consistent with the mere effort account, which would predict that participants subject to evaluation should be motivated to press the key quickly so as to perform as well as possible.

This analysis also produced a significant main effect for type of saccade. Participants exhibited faster adjusted reaction times on correct saccade trials than on corrective saccade trials, $F(1, 52) = 8.81$, $p = .005$, $d = 0.82$. Jamieson and Harkins (2007) also found this effect and argued that it was a result of the fact that the eyes are more likely to arrive at the target site prior to target presentation following correct saccades (no reflexive saccade) than following corrective saccades (after reflexive saccade), and this early arrival may confer some advantage in response preparation.

Prosaccade task. Although there was no significant difference between evaluation participants and no-evaluation participants in their latency to launch a prosaccade, participants in the evaluation condition pressed the response key an average of 76.83 ms faster than did participants in the no-evaluation condition. This response advantage cannot be attributed to eye movement because evaluation and no-evaluation participants arrived at the target location at the same time, on average over 200 ms before the target even appeared. These findings suggest that participants saw the target at the same time, at the 400 ms mark. Thus, it was the motivation to press the key to make the response that produced the reaction time difference between the conditions on the prosaccade trials. This finding is also consistent with the notion that participants in the evaluation condition are motivated to perform well.

Discussion

The overall pattern of performance shows that participants subject to evaluation reported target orientation more quickly than did no-evaluation participants, without any sacrifice in accuracy. Of course, these findings are consistent with the predictions of both the focus of attention and the mere effort accounts. We are able to distinguish between these accounts only when we examine the processes that culminate in the terminal performance measures. In each case, the eye tracking measures are consistent with predictions made by the mere effort account and not with those of the focus of attention account. Participants subject to evaluation made more, not fewer, reflexive saccades. However, they also launched correct and corrective saccades faster than did no-evaluation participants, as well as produced faster adjusted reaction times (times adjusted for the time of arrival of the participants' eyes at the target area).

As a result, even though evaluation participants looked the wrong way more often than did no-evaluation participants, evaluation participants ended up outperforming no-evaluation participants. A closer look at the eye tracking data shows exactly how this happened. As shown in Table 1, on 29.33% of the trials, both participants who were subject to evaluation and those who were not looked in the incorrect direction, toward the cue. On this subset of trials, evaluation participants generated corrective saccades more quickly than did no-evaluation participants. On 53.10% of the trials, both evaluation and no-evaluation participants were able to inhibit the prepotent response and did not look toward the cue. On this subset of trials, the evaluation participants also launched correct saccades more quickly than did no-evaluation participants. Thus, on 82.43% of the trials, evaluation participants launched saccades toward the target more quickly than did no-evaluation participants.

On the remaining 17.57% of the trials, participants subject to evaluation launched reflexive saccades followed by corrective saccades, whereas no-evaluation participants were able to inhibit this response and launched correct saccades. On this subset of trials, it took evaluation participants 356.03 ms to launch a corrective saccade, whereas no-evaluation participants took 333.36 ms to launch correct saccades. In addition, participants in the evaluation condition had to move their eyes twice as far (22°) as no-evaluation participants (11°) to see the target because evaluation participants started their corrective saccades at the cue location, not at the center of the screen. As a result, the eyes of the evaluation participants arrived at the target area on average 38.02 ms after those of the no-evaluation participants. However, the terminal reaction times of evaluation participants on trials when they launched corrective saccades were still faster (M = 401.80 ms, SD = 84.65 ms) than the terminal reaction times of participants in the no-evaluation condition on trials when participants launched correct saccades (M = 512.79 ms, SD = 137.23 ms), $F(1, 55) = 52.16$, $p < .001$, $d = 1.95$. The analysis for reaction times adjusted for arrival time shows that this advantage for the evaluation participants is a direct result of the fact that evaluation participants were more motivated to press the key as quickly as possible than were the no-evaluation participants, which more than made up for their late arrival at the target area.

Table 1
Percentage of Correct and Corrective Antisaccade Trials by Condition

	Correct		Corrective	
Condition	*M*	*SD*	*M*	*SD*
Experimenter evaluation	53.10	21.36	46.90	25.81
No experimenter evaluation	70.67	30.46	29.38	23.38

Note. Correct saccades refer to eye movements directed to the target location, whereas corrective saccades refer to eye movements launched to the target location that were preceded by reflexive saccades made to the cue. These numbers represent the percentage of saccades made out of all valid saccade trials.

Thus, on 82.43% of the trials, the eyes of the evaluation participants arrived at the target area before those of the no-evaluation participants, and as shown by the adjusted reaction time analysis, they also responded to the target more quickly. On the remaining 17.57% of the trials, the eyes of the evaluation participants arrived at the target area after those of the no-evaluation participants, but their motivation to press the key made up for their late arrival.

These findings do not support Muller and Butera's (2007) focus of attention account. Instead of focusing on the central cue (the target) the participants subject to evaluation looked toward the peripheral cue (the box) more, not less, than participants in the no-evaluation condition. The focus of attention account also cannot account for the motivated behavior reflected in the faster saccade launches for correct and corrective saccades, nor can it account for the faster adjusted reaction times. These findings are consistent with the mere effort account and replicate the pattern of findings that Jamieson and Harkins (2007) report in their research on stereotype threat.

However, Jamieson and Harkins (2007; Experiment 1) found that stereotype threat participants performed more poorly (lower accuracy) than did no stereotype threat participants on the antisaccade task at a 150 ms display time. It took a 250 ms display time for the stereotype threat manipulation to produce the same pattern of findings as was produced by the potential for evaluation at 150 ms. We suggest that this difference could result from the fact that Jamieson and Harkins's (2007) stereotype threat manipulation is more potent than is the evaluation manipulation. Consistent with this possibility, stereotype threat participants in Jamieson and Harkins's (2007) research launched reflexive saccades on 59.90% of the trials, whereas participants subject to evaluation did so on only 46.90% of the trials. Thus, stereotype threat participants generated reflexive saccades on 13% more trials than did evaluation participants. At a display time of 150 ms, Jamieson and Harkins (2007) found that stereotype threat participants made errors on 15.70% of antisaccade trials, whereas evaluation participants in the current experiment made errors on only 4.40% of trials at the same display time. Because evaluation participants generated fewer reflexive saccades, it is not surprising that they responded correctly to target orientation, on average, 11.33% more than stereotype threat participants. It is also interesting to note that this 11.33% difference in accuracy approximates the raw difference in the incidence of reflexive responding between threat and evaluation participants (13%).

General Discussion

Five research traditions (e.g., social loafing, goal setting, social facilitation, achievement goal, and intrinsic motivation/creativity) in psychology have offered four different explanations for the effect of the potential for evaluation on complex task performance: processing interference, withdrawal of effort, restricted focus of attention, and drive. On the basis of a molecular analysis of performance on the RAT, Harkins (2006) has proposed another account, mere effort. In this account, he argued that the potential for evaluation motivates participants to perform well, which potentiates the prepotent response. If the prepotent response is correct, performance is facilitated. If the prepotent response is incorrect, and participants do not know, or lack the knowledge or time required for correction, performance is debilitated. In the current research, we used three new tasks (e.g., solving anagrams, the Stroop Color–Word Task, and the antisaccade task) to pit the explanatory power of this mere effort account against the accounts provided by the other explanations.

Experiment 1 allowed us to test mere effort against Bond's (1982) processing interference account and the withdrawal of effort explanation. Research on anagram solution shows that participants attempt to solve these problems by first trying letters in the first position, and because many more words begin with consonants than with vowels, they have a prepotent tendency to begin with consonants (e.g., i.e., Mendelson, 1976; Witte & Freund, 2001). In the mere effort account, it is argued that this prepotent response tendency will be enhanced when there is the potential for evaluation. Thus, the mere effort hypothesis predicts that evaluation participants will be more successful solving anagrams when the solution words begin with consonants but less successful solving anagrams when the solution words begin with vowels than their no-evaluation counterparts.

Word frequency has also been shown to affect the solubility of anagrams: anagrams made from high frequency words are solved more easily than are anagrams made from words of low frequency (e.g., Witte & Freund, 2001). However, this effect is produced through the action of the activation system. As participants begin trying letters (mostly consonants) in the first position, words that begin with those letters gain activation. This activation will add to the resting level of activation for the words, which is higher for high frequency words than for low, making high frequency words easier to solve. As a result, the mere effort explanation would not predict an interaction between evaluation potential and word frequency. Whether the solutions to anagrams are words of high frequency or low frequency, solvers will still tend to try consonants in the first position, and this prepotent response should be even more likely to be made by participants subject to evaluation than by those who are not, yielding only a main effect for word frequency.

In contrast, the withdrawal of effort explanation and Bond's (1982) processing interference account predict an interaction between evaluation potential and other manipulations of difficulty because in these accounts it is argued that what should matter is the potential for evaluation and the experience of difficulty, not the source of the difficulty. However, Bond's (1982) account would predict this outcome only at the level of the aggregate experience, not at the level of the item. That is, he argued that it is the overall experience of difficulty that leads participants to be concerned about how well they are performing rather than the item level of performance predicted by the mere effort and drive/evaluation apprehension accounts.

Experiment 1 was run in two versions to test between these accounts. In one version, initial letter (vowel vs. consonant) was a between-subjects factor and word frequency was a within-subjects factor, whereas in the other version, it was the reverse. To support Bond's (1982) account (or any of the processing interference/withdrawal of effort accounts that would make performance predictions at the aggregate level), we should have found interactions between evaluation and the variable that was manipulated as a between-subjects factor. If these accounts predict item-level effects, we should have found interactions between evaluation and word frequency and evaluation and initial letter, whether the manipulations were within subjects or between subjects. Instead, consistent with the mere effort account, we found a main effect for frequency and an initial letter by evaluation interaction in both versions of the experiment. Thus, the findings for the anagram experiments are consistent with the mere effort account and do not support the explanations offered by the withdrawal of effort explanation or by Bond's processing account.

Experiments 2 and 3, in which the Stroop Color–Word Task was used, allowed us to contrast the mere effort explanation against the focus of attention and the drive/evaluation apprehension accounts. Huguet et al. (1999; see also Huguet et al., 2004) found that social presence enhanced performance on the Stroop and argued that this facilitation was a result of the fact that social presence reduced the range of cues used by the participants. That is, participants saw less of the word, and so, it interfered less with their responses. Huguet et al. (1999) contrasted this focus of attention explanation with Zajonc's (1965) drive theory, which predicts that participants exposed to the color word stimuli will emit the dominant response, reading the word, resulting in debilitation, not facilitation, of performance. Cottrell (1972) made the same prediction but argued that the drive is produced by evaluation apprehension, not mere presence. Mere effort also makes this prediction: Reading the color word is the prepotent, but incorrect response, and the motivation produced by the potential for evaluation should potentiate this incorrect response. However, to observe this debilitation, in the mere effort account, it would be argued that one must use a brief response window. The long response window used by Huguet et al. (1999; Huguet et al., 2004), provides participants subject to evaluation sufficient time to inhibit the incorrect response and still produce the correct response more quickly than do their no-evaluation counterparts.

Thus, the mere effort hypothesis predicts that when time to respond is limited, participants subject to experimenter evaluation will make reliably more errors than will participants who are not, but at longer display times, the potential for evaluation should lead to faster but equally accurate responses. Consistent with the mere effort account, when a 2 s response window was used (Experiment 2), participants subject to experimenter evaluation responded more quickly than did no experimenter evaluation participants without any sacrifice in accuracy. And when brief response windows were used (Exp. 3), participants subject to evaluation made more errors than did no-evaluation participants. Thus, inconsistent with the drive/evaluation apprehension account, at the long display time, evaluation participants outperformed no-evaluation participants, and inconsistent with the focus of attention account, at the short

display time, no-evaluation participants outperformed evaluation participants.

In Experiment 4, we used the antisaccade task to test Muller and Butera's (2007) elaboration of Huguet et al.'s (1999) focus of attention account. Muller and Butera (2007) argued that self-evaluation threat produces ruminations that consume attentional resources that would otherwise be devoted to processing peripheral cues, resulting in focused attention on central cues. However, on the antisaccade task we found that participants subject to evaluation made more, not fewer, reflexive saccades than no-evaluation participants. That is, evaluation participants looked toward the peripheral cue more, not less, than their no-evaluation counterparts. In addition, we found that participants subject to evaluation launched correct and corrective saccades faster than did participants in the no-evaluation condition and had faster adjusted reaction times than did these participants. Each of these effects is inconsistent with the focus of attention explanation; however, the observed pattern of performance is consistent with mere effort.

It should also be noted that the mediating process that Muller and Butera (2007) invoked could lead to a different set of predictions for performance on the antisaccade task. That is, Muller and Butera (2007) argued that self-evaluation threat leads participants to ruminate about the discrepancy between their performance and the participants' standards, which takes up attentional capacity, leading to a restricted focus of attention. Similarly, Schmader and Johns (2003) have argued that when under stereotype threat, participants expend cognitive resources that could be devoted to task performance on processing information resulting from the activation of the negative stereotype. Thus, in each case, participants are using processing capacity to ruminate about their task performance. However, instead of leading to reduced focus of attention, Schmader and Johns (2003) argued that the reduction in working memory capacity directly produces the performance debilitation reported in the stereotype threat literature (e.g., Cadinu, Maass, Rosabianca, & Kiesner, 2005; Croizet et al., 2004).

More specifically, Schmader and Johns (2003) argued that the executive attention component (central executive) of working memory (Engle, 2001; 2002) is impaired by the ruminations. The central executive is essential for effective performance on inhibition tasks, like the antisaccade task. Thus, if evaluation potential produces ruminations, which interfere with working memory, participants subject to evaluation should produce more reflexive saccades than should controls because participants subject to evaluation have less ability to inhibit their tendency to look at the cue. They should also launch correct and corrective saccades more slowly than should control participants because the capacity to launch these eye movements also requires the central executive (Kane et al., 2001; Roberts et al., 1994; Stuyven et al., 2000; Unsworth et al., 2004). However, as described previously, although in the current research it was found that participants subject to evaluation generated more reflexive saccades than did controls, participants subject to evaluation also launched correct and corrective saccades faster than did control participants and had faster adjusted reaction times. Each of these effects indicates that participants' central executive processes were not impaired by the potential for evaluation. Thus, the findings of Experiment 4 are consistent with the mere effort account, but the findings are consistent neither with Muller and Butera's (2007) focus of attention account nor with alternative predictions that would follow from a working memory explanation.

In addition to providing tests of focus of attention and processing interference accounts, the use of the eye tracker on the antisaccade task allowed us to examine the mechanisms that form the core of the mere effort account: response potentiation and the correction process. This analysis first shows that evaluation participants made reflexive saccades on 46.90% of antisaccade trials, whereas no-evaluation participants generated these eye movements on only 29.33% of trials. At this point, participants either launched corrective (after an incorrect reflexive saccade to the cue) or correct (no reflexive saccade) saccades to the target location. Because both are volitional (i.e., endogenous), motivation should reduce the latency to launch each type of saccade. Consistent with this notion, evaluation participants launched corrective saccades in 356.03 ms, whereas controls made these eye movements in 405.64 ms. Evaluation participants also launched correct saccades faster than did no-evaluation participants (283.19 ms vs. 333.36 ms).

Taking into account the frequency and the latency to launch each type of eye movement as well as travel time, on average, evaluation participants' eyes arrived at the target site at 376.11 ms, whereas no-evaluation participants' eyes arrived at 406.29 ms, a 30.18 ms advantage. However, on a number of these trials, the participants' eyes arrived at the target site before the target had appeared. Taking into account early arrival trials, the motivation to generate volitional eye movements actually accounts for 25.17 ms of advantage in terminal reaction time for evaluation participants over controls. This 25.17 ms advantage, however, does not account for the overall terminal reaction time advantage (150.86 ms) for evaluation participants. The remainder is accounted for by the adjusted key press analysis.

Adjusted reaction time indexes how motivated participants are to press the key, controlling for when their eyes arrive at the target site. This measure showed that once their eyes arrived at the target site, evaluation participants took 359.79 ms to respond to the target's orientation, whereas no-evaluation participants took 485.48, a 125.69 ms advantage for evaluation participants. The contribution of the adjusted reaction times (125.69 ms) plus the contribution of the eye movements (25.17 ms) fully accounts for the finding that the terminal reaction time of participants subject to evaluation (383.79 ms) is 150.69 ms faster than the time for no-evaluation participants (534.65 ms). Thus, we show that the proportion of reflexive saccades, correct and corrective saccade launch latencies, and adjusted reaction times account for the mean difference in the dependent variable (terminal reaction time) produced by the independent variable (evaluation).

This line of research also makes a compelling case for the view that we must adopt a more sophisticated approach to the tasks that we use in our research. Certainly, knowing that participants find a task difficult is not enough to be able to predict the effect that the potential for evaluation will have on performance. For example, our research shows that one can take two difficult tasks, and on one the potential for evaluation improves performance (e.g., anagrams made from low frequency words that all begin with consonants), whereas on another it debilitates performance (e.g., anagrams made from words that begin with vowels). One can also find that evaluation debilitates performance on tasks that would appear to be simple. For example, on the Stroop, despite the fact that overall, participants perform extremely well, when participants subject to

evaluation are required to respond in a brief interval, their performance is worse than that of participants who are not subject to this evaluation. Thus, as these examples show, knowing that a task is simple or difficult does not allow prediction of the effect that evaluation will have on performance.

Finally, effective task performance, whether in educational or work settings, is essential to the success of individuals as well as of our society. However, it is also in settings like these that variables like the potential for evaluation is likely to arouse performance concerns. Knowing the specific process(es) that mediate(s) the effects of evaluation on performance is required for the design of effective intervention strategies. For example, if we believe that performance is debilitated simply because participants withdraw effort, we could try to persuade them not to do so. However, our research suggests that this intervention could have an effect opposite to that intended. It is a high level of motivation that may be producing the problem in the first place. Thus, rather than attempting to motivate individuals to try harder or to improve their cognitive processing, interventions may seek to take advantage of the individuals' heightened motivational state by helping them to direct their efforts more effectively. For example, Harkins (2006) found that instructing participants subject to evaluation to simply register the words in a RAT triad and to then wait for the answer to pop up significantly improved performance. Thus, in addition to advancing our understanding of an issue that dates from the birth of experimental social psychology (Triplett, 1898), this research may also provide the basis for the design of intervention strategies that allow individuals to maximize their performance.

References

Amabile, T. (1979). Effects of external evaluation on artistic creativity. *Journal* of *Personality and Social Psychology, 37,* 221–233.

Bach, M., Bouis, D., & Fischer, B. (1983). An accurate and linear infrared oculometer. *Journal of Neuroscience Methods, 9,* 9–14.

Baron, R. (1986). Distraction-conflict theory: Progress and problems. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 1–40). New York: Academic Press.

Bond, C. (1982). Social facilitation: A self-presentational view. *Journal of Personality and Social Psychology, 42,* 1042–1050.

Cadinu, M., Maas, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? *Psychological Science, 16,* 572–578.

Carver, C., & Scheier, M. (1981). The self-attention-induced feedback loop and social facilitation. *Journal of Experimental Social Psychology, 17,* 545–568.

Cottrell, N. B. (1968). Performance in the presence of other human beings: Mere presence, audience, and affiliation effects. In E. C. Simmell, R. A. Hoppe, & G. A. Milton (Eds.), *Social facilitation and imitation behavior* (pp. 28–37). Boston: Allyn & Bacon.

Cottrell, N. B. (1972). Social facilitation. In N. B. McClintock (Ed.), *Experimental social psychology* (pp. 185–236). New York: Holt, Rinehart, & Winston.

Crevits, L., & Vandieraendonck, A. (2005). Gap reflex in reflexive and intentional prosaccades. *Neuropsychobiology, 51,* 39–44.

Croizet, J. C., Després, G., Gauzins, M. E., Huguet, P., Leyens, J. P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin, 30,* 721–731.

Elliott, A., Shell, M., Henry, K., & Maier, M. (2005). Achievement goals, performance contingencies, and performance attainment: An experimental test. *Journal of Educational Psychology, 97,* 630–640.

Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, DC: American Psychological Association.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23.

Ford, K. A., Goltz, H. C., Brown, M. R. G., & Everling, S. (2005). Neural processes associated with antisaccade task performance investigated with event-related fMRI. *Journal of Neurophysiology, 94,* 429–440.

Geen, R. (1989). Alternative conceptions of social facilitation. In P. Paulus (Ed.), *Psychology of group influence* (pp. 15–51). Hillsdale, NJ: Erlbaum.

Hallet, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research, 18,* 1279–1296.

Harkins, S. (2001). The three-variable model: From Occam's razor to the black box. In S. Harkins (Ed.), *Multiple perspectives on the effects of evaluation on performance: Toward an integration* (pp. 99–131). Kluwer Academic: Norwell, MA.

Harkins, S. (2006). Mere effort as the mediator of the evaluation–performance relationship. *Journal of Personality and Social Psychology, 91,* 436–455.

Hennessey, B. A. (2001). The social psychology of creativity: Effects of evaluation on intrinsic motivation and creativity of performance. In S. Harkins (Ed*.), Multiple perspectives on the effects of evaluation on performance: Toward an integration* (pp. 99–131). Kluwer Academic: Norwell, MA.

Hochman, S. (1967). The effects of stress on Stroop color–word performance. *Psychonomic Science, 9,* 475–476.

Hochman, S. (1969). Stress and response competition in children's color–word performance. *Perceptual and Motor Skills, 28,* 115–118.

Houston, B. K. (1969). Noise, task difficulty, and Stroop color–word performance. *Journal of Experimental Psychology, 82,* 403–404.

Houston, B. K., & Jones, T. (1967). Distraction and Stroop color–word performance. *Journal of Experimental Psychology, 74,* 54–56.

Huguet, P., Dumas, F., & Monteil, J. (2004). Competing for a desired reward in the Stroop task: When attentional control is unconscious but effective versus conscious but ineffective. *Canadian Journal of Experimental Psychology, 58,* 153–167.

Huguet, P., Galvaing, M., Monteil, J., & Dumas, F. (1999). Social presence effects in the Stroop task: Further evidence for an attentional view of social facilitation. *Journal of Personality and Social Psychology, 77,* 1011–1025.

Jackson, J., & Williams, K. (1985). Social loafing on difficulty tasks: Working collectively can improve performance. *Journal of Personality and Social Psychology, 49,* 937–942.

Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology, 93,* 544–564.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General, 130,* 169–183.

Kirk, R. (1995). *Experimental design.* Pacific Grove, CA: Brooks/Cole Publishing.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Locke, E., & Latham, G. (1990). *A theory of goal setting and task performance.* Englewood Cliffs, N J: Prentice Hall.

MacKinnon, D., Geiselman, E., & Woodward, J. (1985). The effects of effort on Stroop interference. *Acta Psychologica, 58,* 225–235.

Mayzner, M., & Tresselt, M. (1958). Anagram solution times: A function of letter order and word frequency. *Journal of Experimental Psychology, 56,* 376–379.

Mendelson, G. (1976). A hypothesis approach to the solution of anagrams. *Memory & Cognition, 4,* 637–642.

Montagnini, A., & Chelazzi, L. (2005). The urgency to look: Prompt saccades to the benefit of perception. *Vision Research, 45,* 3391–3401.

Muller, D., Atzeni, T., & Butera, F. (2004). Coaction and upward social comparison reduce the illusory conjunction effect: Support for distraction-conflict theory. *Journal of Experimental Social Psychology, 40,* 659–665.

Muller, D., & Butera, F. (2007). The focusing effect of self-evaluation threat in coaction and social comparison. *Journal of Personality and Social Psychology, 93,* 194–211.

O'Malley, J., & Poplawsky, A. (1971). Noise-induced arousal and breadth of attention. *Perceptual and Motor Skills, 33,* 887–890.

Pallak, M., Pittman, T., Heller, J., & Munson, P. (1975). The effect of arousal on Stroop color–word task performance. *Bulletin of the Psychonomic Society, 6,* 248–250.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Remington, R. W., Johnston, J. C., & Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception and Psychophysics, 51,* 279–290.

Roberts, R., Hager, L., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology; General, 123,* 374–393.

Sarason, I., Pierce, G., & Sarason, B. (1996). Domains of cognitive interference. In I. Weiner (Ed.), *Cognitive interference* (pp. 139–152). Mahwah, N J: Erlbaum.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452.

Sereno, A. B. (1992). Programming saccades: The role of attention. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 89–107). New York: Springer.

Stroop, J. R. (1935). Studies of interference in serial-verbal reaction. *Journal of Experimental Psychology, 18,* 643–662.

Stuyven, E., Van der Goten, K., Vandieraendonck, A., Claeys, K., & Crevits, L. (2000). The effect of cognitive load on saccadic eye movements. *Acta Psychologica, 104,* 69–85.

Treisman, A. (1988). Features and objects: The Fourteenth Bartlett Memorial Lecture. *The Quarterly Journal of Experimental Psychology, 40,* 201–237.

Treisman, A., & Paterson, R. (1984). Emergent features, attention, and object perception. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 12–31.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107–141.

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology, 9,* 507–533.

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1302–1321.

Witte, K., & Freund, J. (2001). Single-letter retrieval cues for anagram solution. *Journal of General Psychology, 128,* 315–328.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Performance and Perception, 10,* 601–621.

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Performance and Perception, 16,* 121–134.

Zajonc, R. (1965, July 16). Social facilitation. *Science, 149,* 269–274.

Received January 6, 2006
Revision received May 21, 2008
Accepted May 22, 2008 ■