Journal of Experimental Psychology: General
2007, Vol. 136, No. 2, 256–276

Copyright 2007 by the American Psychological Association
0096-3445/07/$12.00 DOI: 10.1037/0096-3445.136.2.256

Stereotype Threat and Working Memory: Mechanisms, Alleviation, and Spillover

Sian L. Beilock
University of Chicago

Robert J. Rydell
University of California, Santa Barbara

Allen R. McConnell
Miami University

Stereotype threat (ST) occurs when the awareness of a negative stereotype about a social group in a particular domain produces suboptimal performance by members of that group. Although ST has been repeatedly demonstrated, far less is known about how its effects are realized. Using mathematical problem solving as a test bed, the authors demonstrate in 5 experiments that ST harms math problems that rely heavily on working memory resources—especially phonological aspects of this system. Moreover, by capitalizing on an understanding of the cognitive mechanisms by which ST exerts its impact, the authors show (a) how ST can be alleviated (e.g., by heavily practicing once-susceptible math problems such that they are retrieved directly from long-term memory rather than computed via a working-memory-intensive algorithm) and (b) when it will spill over onto subsequent tasks unrelated to the stereotype in question but dependent on the same cognitive resources that stereotype threat also uses. The current work extends the knowledge of the causal mechanisms of stereotype threat and demonstrates how its effects can be attenuated and propagated.

Keywords: stereotype threat, working memory, stereotyping, worries, choking under pressure

Theories of stereotype threat suggest that introducing a negative stereotype about a social group in a particular domain can reduce the quality of task performance exhibited by group members (Steele, 1997). When negative group stereotypes are activated in performance situations, African Americans perform poorly on cognitive tasks reputed to assess intelligence (Steele & Aronson, 1995), women perform at a less-than-optimal level on math problems for which they have been told gender differences exist (Spencer, Steele, & Quinn, 1999), and Whites perform poorly on athletic tasks that are purportedly diagnostic of athletic ability rather than athletic intelligence (Beilock, Jellison, Rydell, McConnell, & Carr, 2006; Stone, Lynch, Sjomeling, & Darley, 1999).

Causal Mechanisms of Stereotype Threat

Although stereotype threat has been demonstrated across a wide range of social groups and task types (see Wheeler & Petty, 2001), only recently has its underlying causal mechanisms received attention. Schmader and Johns (2003) argue that stereotype threat interferes with performance by reducing the working memory capacity that individuals need to perform a task successfully. Working memory can be thought of as a short-term memory system involved in the control, regulation, and active maintenance of a limited amount of information with immediate relevance to the task at hand (Miyake & Shah, 1999a). If the capacity of the working memory system to oversee task-relevant information is disrupted, performance may suffer.

Schmader and Johns (2003) tested the relation between working memory and stereotype threat by activating negative, self-relevant stereotypes in women (highlighting gender differences in quantitative ability) and Latinos (stressing ethnic group intelligence differences) and then measuring the working memory capacity of stereotyped group members. Working memory was significantly lower for both women and Latinos after receiving the stereotype threat manipulation in comparison to individuals who did not receive the negative stereotypes. In a follow-up experiment, women completed a working memory task and a difficult math task under control conditions or following the activation of a stereotype regarding gender differences in quantitative ability. Stereotype threat led to poorer math performance, and working memory capacity mediated this relationship.

The above findings support a causal role of working memory in reduced math performance under stereotype threat. Yet what exactly stereotype threat does to this cognitive system to produce suboptimal performance in tasks such as math problem solving remains unclear. An understanding of the locus of stereotype threat effects is not only important for establishing a fuller theoretical account of this phenomenon, but such knowledge will (a) inform the development of training regimens and testing situations de-

Sian L. Beilock, Department of Psychology, University of Chicago; Robert J. Rydell, Department of Psychology, University of California, Santa Barbara; Allen R. McConnell, Department of Psychology, Miami University.

This research was supported by Institute of Education Sciences Grant R305H050004 to Sian L. Beilock and National Science Foundation Grant BCS-0601148 to Sian L. Beilock and Allen R. McConnell.

Correspondence concerning this article should be addressed to Sian L. Beilock, Department of Psychology, 5848 South University Avenue, University of Chicago, Chicago, IL 60637. E-mail: beilock@uchicago.edu

signed to ameliorate these unwanted performance decrements and (b) shed light on when stereotype threat effects will persist—even in domains not necessarily implicated by the negative stereotype in question.

Performance Pressure and Test Anxiety

One approach to elucidating the causal mechanisms of stereotype threat is to examine other literatures exploring unwanted performance decrements in cognitively demanding tasks. In the domain of mathematical problem solving, Ashcraft and Kirk (2001) have proposed that anxiety about math computations drains the working memory capacity that might otherwise be available for math performance by inducing intrusive thoughts and worries that compete with the on-going cognitive task (for a more general theory of such anxiety-induced worries, see Eysenck & Calvo, 1992). Beilock and colleagues (Beilock & Carr, 2005; Beilock, Kulp, Holt, & Carr, 2004) have come to similar conclusions regarding the impact of situation-induced feelings of performance pressure on math task execution. Despite the implication of working memory consumption as a source of failure in these literatures however, how such failure actually occurs has not been adequately addressed. For example, there is little evidence that verbal worries cause performance decrements in high-pressure situations or in highly math-anxious individuals. Rather, existence of worries is often inferred by a performance drop on working memory demanding tasks (e.g., see Ashcraft & Kirk, 2001).

Moreover, because stereotype threat, performance pressure, and math anxiety remain largely separate research areas to date (cf. Smith & Johnson, 2006), the relation between these phenomena is not well understood. Investigating how stereotype threat operates is important then, not only given the dearth of work identifying its cognitive processes but also for the development of comprehensive theories of skill failure that simultaneously take into account social and cognitive factors related to both the performer and the task being performed.

Multicomponent Model of Working Memory

We looked to theories addressing working memory's organization in order to tackle the question of how stereotype threat might impact working memory in tasks such as math problem solving. Although there are a number of prominent working memory models that differ on both structural and functional dimensions (see Miyake & Shah, 1999b), research examining working memory in the context of mathematical computation has most often used Baddeley's (1986; Baddeley & Logie, 1999) multicomponent model as a guide. Baddeley's original multicomponent model, in which different subsystems are thought to be devoted to different types of information, had three major components—a limited-capacity central executive, a phonological loop for storing verbal information, and a visual–spatial sketchpad for storing visual images. A fourth component has also been added—a multimodal episodic buffer that serves to bind information from the phonological loop, the visual–spatial sketchpad, and long-term memory into a unitary episodic representation (Baddeley, 2000).

Because the verbal–visuospatial distinction has received a large amount of support in the human working memory literature (DeStefano & LeFevre, 2004; Gray, 2001), conceptualizing differences in math task demands in terms of verbal and visuospatial processing requirements provides a useful approach for examining how stereotype threat effects occur. Nonetheless, there is debate concerning whether working memory should be viewed primarily as a domain-general unitary system involved in executive-attention function (Cowan, 1999; Engle, Kane, & Tuholski, 1999; Lovett, Reder, & Lebiere, 1999) or as a domain-specific system consisting of specialized components that handle specific types of information (Baddeley, 1986; Baddeley & Logie, 1999; Friedman & Miyake, 2000; Shah & Miyake, 1996). Individuals who argue for a domain-specific view do not deny that domain-general components exist (Miyake, 2001). Furthermore, models that support a domain-general view of working memory find evidence for, in addition to domain-general control processes, domain-specific verbal and visuospatial processes (Engle et al., 1999; Kane et al., 2004). Thus, to the extent that different types of math problems share domain-general processing demands but can be differentiated in terms of the specific demands they make on verbal and visuospatial resources, insight into how stereotype threat harms the working memory system can be realized.

Returning to Stereotype Threat

Steele, Spencer, & Aronson (2002) have suggested that stereotype threat is accompanied by "concerns about how one will be perceived, doubts about one's ability, thoughts about the stereotype. . ." (p. 392). Recent work by Cadinu, Maass, Rosabianca, and Kiesner (2005) supports this idea. Women performing math problems after being told that gender differences in math exist (i.e., stereotype threat) not only performed worse than a control group of women but also reported having more negative math-related thoughts than women who did not receive this information.

If situation-induced worries underlie performance decrements in stereotype threat situations, how might such thoughts exert their impact on the working memory system? One possibility is that worries and verbal ruminations occupy central executive resources needed for integrating and monitoring the step-by-step processes of performance (Ashcraft & Kirk, 2001). Some support for this notion comes from the finding that stimulus-independent thoughts unrelated to immediate sensory input (e.g., daydreaming) tax central executive resources (Teasdale et al., 1995). It should be noted however, that stimulus-independent thoughts are characterized as having no relation to the task at hand (Christoff, Ream, & Gabrieli, 2004) and thus their intensity, composition, and cognitive underpinnings may not necessarily be the same as task-related worries about the situation and its consequences.

A second possibility is that pressure-induced worries rely more heavily on the phonological aspect of working memory, which is thought to support inner speech and thinking in the service of complex cognitive activities (Carlson, 1997; Miyake & Shah, 1999a). Research in the test anxiety literature indicating that the representation and rehearsal of unwanted thoughts and worries impacts the phonological resources of working memory supports this notion (Darke, 1988; Ikeda, Iwanaga, & Seiwa, 1996; Markham & Darke, 1991; Rapee, 1993). These thoughts and worries may also have some impact on central executive resources. Nonetheless, to the extent that such thoughts rely on phonological resources as well, two types of math problems that are equally dependent on central executive processes (e.g., because they re-

quire the same algorithmic computation; Baddeley & Logie, 1999), but are differentially dependent on phonological resources (e.g., because the maintenance and rehearsal of intermediate steps are represented in different forms) may show differential outcomes under stereotype threat.

The current work draws upon research demonstrating that the orientation of a presented math problem can alter the working memory resources it relies on (Trbovich & LeFevre, 2003) to explore how stereotype threat impacts working memory. Knowledge regarding the locus of stereotype threat effects within the working memory system is then used to (a) design training regimens to alleviate unwanted performance decrements and (b) predict when such effects will persist—even when the task being performed is no longer related to the stereotype in question.

Experiment Overview

We began by demonstrating stereotype threat. Specifically, we examined whether women, who received the information that they were participating in research investigating why men are generally better at math than women, would perform worse on a math problem-solving task than would women who did not receive this information.

In Experiment 2, we set the stage to examine the types of problems most susceptible to stereotype threat. Individuals judged the validity of horizontally oriented and vertically oriented math problems (shown to be more and less reliant on phonological resources, respectively; Trbovich & LeFevre, 2003) under both a single-task and a phonological load condition. Because of the hypothesized role of verbal thoughts and worries in stereotype threat (Cadinu et al., 2005; Steele et al., 2002) and the impact such thoughts may have on verbal working memory resources, our goal was to identify problems that depend heavily on verbal resources in order to test whether stereotype threat is most strongly revealed for such problems.

In Experiment 3, women performed either horizontally presented or vertically presented math problems in both a baseline and a subsequent stereotype threat condition and reported their thoughts and worries under stereotype threat. If stereotype threat taxes verbal working memory resources, then those problems that rely most heavily on this capacity (i.e., horizontally presented problems) should be most likely to fail. As a preview, this is exactly what was found. In a follow-up experiment (Experiment 3B), women performed horizontal or vertical math problems in a no stereotype threat control condition and reported the thoughts they had while performing the math problems. Women in this control condition performed at a consistently high level—regardless of math problem orientation—and reported worrying significantly less about the situation and its consequences than did individuals under stereotype threat.

Experiment 4 explored ways to mitigate stereotype threat. Specifically, if the performance of horizontal math problems is harmed because stereotype threat impinges on the processing resources needed for successful execution, then making such problems less reliant on working memory should alleviate the impact of a negative performance stereotype.

Finally, Experiment 5 explored a novel implication of the hypothesis that stereotype threat harms math task performance via the consumption of working memory resources—and especially verbal resources. Individuals performed horizontal math problems under stereotype threat followed by either a verbal or spatial computerized two-back working memory task. These tasks were matched for difficulty and appearance, differing most substantially in their reliance on either verbal or spatial working memory processes (Gray, 2001). Borrowing logic from the finding that depletion of resources in one task domain can carry over and impact performance on another task (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998), if stereotype threat selectively impinges on verbal processing resources (e.g., via worries about the situation and its consequences) and this working memory consumption does not immediately subside when performance on the stereotyped task is finished, then individuals should perform poorer on a verbal (relative to a spatial) two-back task following stereotype threat. In essence, stereotype threat may "spill over" onto other tasks that use the same processing resources but that are not implicated by the negative stereotype. Not only would this be the first demonstration that a cultural stereotype can adversely affect performance in domains unrelated to the stereotype in question, but it would also have important applied implications (e.g., the ordering of quantitative vs. verbal sections on standardized tests may have unanticipated performance consequences).

We used *modular arithmetic* (Gauss, 1801, as cited in Bogomolny, 1996) as our math task. The object of modular arithmetic (MA) is to judge the validity of problems such as *51 = 19 (mod 4)*. To do this, the middle number is subtracted from the first number (i.e., 51 − 19), and then this difference is divided by the last number (i.e., 32 ÷ 4). If the dividend is a whole number, the problem is "true." MA is an advantageous math task because its working memory demands can be easily manipulated. Working memory demand was determined by whether the first step of the MA problem involved numbers greater than 10 and a borrow operation (e.g., *45 = 27 (mod 4)*). Larger numbers and borrow operations involve longer sequences of steps and require maintenance of more intermediate products, placing greater demands on working memory (Ashcraft, 1992; Ashcraft & Kirk, 2001; Beilock & Carr, 2005; DeStefano & LeFevre, 2004). If stereotype threat exerts its impact by co-opting working memory resources on which MA problems rely, then performance on higher working memory demanding problems should be most likely to fail.

Moreover, across all experiments, half of the MA equations presented to participants were "true," and the rest were "false." Additionally, each "true" problem had a "false" correlate that only differed as a function of the number involved in the mod statement. For example, if the "true" problem *51 = 19 (mod 4)* was presented, then a "false" correlate problem *51 = 19 (mod 3)* was also presented at some point in the same problem block. This pairing was designed to equate the "true" and "false" problems as much as possible in terms of the specific numbers used in each equation. Finally, to equate the difficulty of the horizontal and vertical problems within each experiment, the problems within each orientation were counterbalanced across participants (e.g., horizontal problems presented to one participant were presented as vertical problems to another).

Participants were undergraduate students. To ensure that all participants demonstrated reasonable performance on the MA task prior to the introduction of any experimental manipulations, only individuals whose problem-solving accuracy was greater than 75% in the practice and baseline blocks were retained as participants.

The first four experiments were run at the same Midwestern university and the fifth experiment at a second Midwestern university.

Experiment 1

Women randomly assigned to a no threat (control) or stereotype threat (ST) group performed horizontal MA problems. Problems were either lower or higher in working memory demands. To the extent that stereotype threat exerts its impact by co-opting verbal working memory resources, horizontally presented math problems that rely heavily on such resources (Trbovich & LeFevre, 2003) should be especially susceptible to failure. Such a result would be consistent with the general finding in academically related cognitive tasks that "stereotype threat effects have been consistently greatest for more difficult tasks" (Steele et al., 2002, p. 391). Moreover, it would also provide a mechanistic advance by demonstrating that the locus of this difficulty effect is dependence on a limited capacity working memory system.

Method

Participants. Thirty-one women participated, each of whom met the aforementioned criteria. Moreover, Steele (1997) suggested that negative stereotypes have little effect on individuals who are not skilled and do not value the domain associated with the stereotype. Thus, to be retained as participants, individuals had to have reported at least moderate levels of math skills and importance of these skills (an average rating greater than 5, the midpoint, of two 9-point math-related questions "I am good at math" and "It is important to me that I am good at math"). Similar identification criteria have been used in previous research to ensure that only individuals most susceptible to stereotype threat serve as participants (Aronson et al., 1999; Spencer et al., 1999).

Seventeen individuals were randomly assigned to the control group and 14 participants to the ST group. The control and ST groups did not differ in terms of their perceptions of their math skill (control: $M = 6.88$, $SE = 0.31$; ST: $M = 7.35$, $SE = 0.27$), $F(1, 29) = 1.29$, $p = .26$, or the importance assigned to this skill (control: $M = 6.94$, $SE = 0.36$; ST: $M = 7.43$, $SE = 0.23$), $F(1, 29) = 1.19$, $p = .28$.

Procedure. Participants completed a consent form informing them that the purpose of the study was to examine how individuals learn a new math skill. Individuals were introduced to MA by written instructions presented on a computer. Participants were instructed to judge the validity of each problem as quickly as possible without sacrificing accuracy, indicating their response using the *T* or *F* keys on a standard keyboard. Participants were instructed to rest their right and left index fingers on the *T* and *F* keys, respectively, throughout the experiment.

Each trial began with a 500-ms fixation point in the center of the screen, which was immediately replaced by an MA problem present until response. The problem was then extinguished and the word *Correct* or *Incorrect* was displayed on the screen for 1,000 ms, providing feedback. The screen then went blank for a 1,000-ms intertrial interval.

Everyone first performed 12 practice problems presented in a different random order to each participant. Four problems were considered low demand as they required a single-digit no borrow subtraction operation (e.g., *7 = 2 [mod5]*). Four high-demand problems required a double-digit borrow subtraction operation (e.g., *43 = 16 [mod3]*). Four filler problems exerting intermediate capacity demands (requiring a double-digit no borrow subtraction operation, *19 = 12 [mod7]*) diminished the contrast between low- and high-demand problems.

All participants then completed two blocks of 24 problems, each consisting of 8 low-demand, 8 high-demand, and 8 fillers. Problems within each block were presented in a different random order and counterbalanced across participants. Problems were presented once.

The first block of problems (baseline) served as an initial performance measure for both the control and ST groups. Immediately preceding the second block of problems (posttest), Control participants read on the computer that the experiment was investigating why some people are better at math than are others. ST participants read that the research was investigating why men are generally better than women at math. Manipulation wording (see Appendix) was adapted from Aronson et al. (1999). Following the posttest, participants were debriefed.

Results

Specific MA problems (and their response times [RTs]) that were not performed at least 65% correct across all participants in the baseline condition were removed from both the baseline and experimental blocks in all experiments to ensure that individual MA problems were not unduly difficult to solve. Three problems were removed from Experiment 1. In addition, to reduce the positive skew of RTs and thus the impact of outliers, RTs were log transformed for the analyses. However, for ease of comprehension, raw means are reported. Finally, 95% confidence intervals were used to assess significance for all simple effects.

Accuracy and corresponding RTs for MA problems to which responses were correct were compared in a 2 (group: control, ST) × 2 (block: baseline, posttest) × 2 (problem working memory demand: low demand, high demand) design, with group as a between-subjects variable.

In terms of accuracy, a significant Group × Block × Problem Demand interaction obtained, $F(1, 29) = 11.18$, $p < .01$, $\eta_p^2 = .28$ (see Figure 1). A 2 (block: baseline, posttest) × 2 (problem demand: low demand, high demand) analysis of variance (ANOVA) for the control group revealed only a main effect of problem difficulty, $F(1, 16) = 15.69$, $p < .01$, $\eta_p^2 = .50$. Not surprisingly, accuracy was higher for the low-demand than for the high-demand problems. The same ANOVA for the ST group revealed a Block × Difficulty interaction, $F(1, 13) = 7.18$, $p < .01$, $\eta_p^2 = .36$. There was no difference between the ST group's low-demand problem performance from the baseline to the posttest. However, high-demand problem accuracy was significantly lower in the posttest ($M = 79.3\%$, $SE = 4.6\%$) as compared with the baseline condition ($M = 89.1\%$, $SE = 3.8\%$; confidence interval [CI]: 81.0%−97.0%; $d = 0.61$).

In terms of RTs, a 2 (group: control, ST) × 2 (block: baseline, posttest) × 2 (problem working memory demand: low demand, high demand) ANOVA revealed main effects of block, $F(1, 29) = 8.33$, $p < .01$, $\eta_p^2 = .22$, in which individuals performed the problems faster over time, and problem demand, $F(1, 29) = 754.5$, $p < .01$, $\eta_p^2 = .96$, in which high-demand problem RTs were

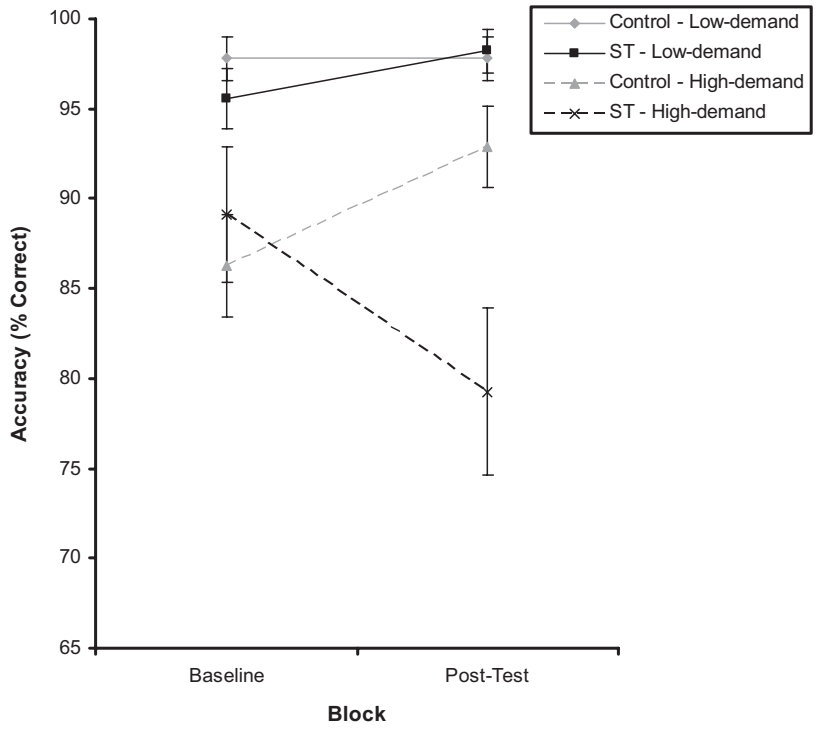

Figure 1. Accuracy (percentage correct) in the baseline and posttest for the low-demand and high-demand horizontal problems for the stereotype threat and control groups in Experiment 1. Error bars represent standard errors.

slower than were low-demand RTs (see Table 1). All other main effects and interactions, including the Group × Block × Problem Demand interaction were not significant (Fs < 1).

Discussion

Participants assigned to a control or ST group performed horizontally presented MA problems that varied as a function of the demands they placed on working memory. Only MA problems heavily dependent on working memory (i.e., horizontal high-demand problems) failed under stereotype threat, suggesting that stereotype threat exerts its impact by co-opting working memory resources needed for the successful execution of such problems. In Experiment 2, we examined whether horizontal high-demand problems are more reliant on phonological resources than other problem types. Such a finding sets the stage for identifying the locus of stereotype threat effects within the working memory system.

Experiment 2

Participants performed horizontally and vertically oriented math problems[1] (see Figure 2) in both a single-task and a phonological load dual-task condition. Although all arithmetic problems involve central executive resources (DeStefano & LeFevre, 2004), Trbovich and LeFevre (2003) demonstrated that math problems presented in a horizontal format depend heavily on phonological resources as well. This is thought to be due, in part, to the selection of solution procedures that require the verbal maintenance of intermediate steps in memory. In contrast, Trbovich and LeFevre found that math problems presented in a vertical format rely more heavily on visuospatial resources as individuals tend to solve vertical problems in a spatial mental work space similar to how such problems are solved on paper. If horizontal problems recruit verbal working memory resources that vertical problems do not, then performance on the horizontal problems should be more negatively impacted by the phonological secondary task than vertical problems. Experiment 2 tested this notion.

Method

Participants. Twenty-four individuals participated in this experiment, each of whom met the criteria outlined in the experiment overview.

Procedure. Participants provided informed consent and were introduced to MA. Individuals first performed a practice block of 8 MA problems (4 vertical, 4 horizontal) presented in a different random order to each participant. Within orientation, half of the problems were high in working memory demands and half were low in working memory demands.

[1] Previous research has demonstrated that the locus of the working memory demands of MA problems occurs most strongly within the subtraction procedure (Beilock & Carr, 2005; Beilock et al., 2004). Thus, horizontal versus vertical orientation was altered in the same portion of the MA problem.

Table 1
Mean Response Times as a Function of Stereotype Threat Condition, Block, and Problem Working Memory Demand in Experiment 1

	Low demand				High demand			
	Baseline		Posttest		Baseline		Posttest	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Control	2,178	164	1,954	106	7,893	422	7,486	445
Stereotype threat	2,114	128	1,894	87	8,133	692	7,150	735

Note. Mean response times are reported in real-time metric (i.e., ms).

Following practice, all participants completed a single-task baseline block consisting of 32 problems (16 horizontal, 16 vertical) presented in a different random order to each participant. As in the practice, half of the problems within each orientation were low-demand problems and half were high-demand problems. Performance feedback was not given in the single-task baseline.

Individuals were next introduced to the phonological secondary task via instructions presented on the computer (see the following section for phonological task details). Participants were informed that they should try to perform both tasks as quickly and accurately as possible, not favoring one task over the other. Individuals performed several problems (half horizontal, half vertical) along with the secondary task to familiarize them with the dual-task procedure.

Next, individuals completed a 32 problem dual-task block presented in a different random order to each participant. As in the single-task baseline, there was no performance feedback. Half of the problems were horizontally oriented and half were vertically oriented. Within these orientations, half were low-demand and half were high-demand problems. Problems were presented once across the experiment. Following the dual-task block, individuals were debriefed.

Phonological secondary task. This task was adapted from Trbovich and LeFevre (2003). Participants solved MA problems while they retained a phonological load in memory. Three pronounceable nonwords consisting of a consonant–vowel–consonant (e.g., *gib, lec, nup*) were presented on the screen for 1,500 ms. The screen then went blank (1 s) so that participants could rehearse the nonwords. Next an MA problem appeared on the screen until participants responded by pressing either the *T* or *F* key. Another nonword then appeared (e.g., *geb*), remaining on screen until participants indicated whether it was the same as any of the nonwords presented prior to the MA problem. In cases in which the second nonword had not been presented previously, it differed from one of the first three nonwords by only one letter. The visual similarity of the second nonword to the nonwords presented prior to the MA problem was designed to ensure that participants were using phonological coding (that discriminates between different phonemes) rather than visual codes (that would not discriminate as successfully among visually similar phonemes) to maintain the letter strings in memory (DeStefano & LeFevre, 2004). Individuals pressed the *T* key if the second nonword was the same as any of the nonwords presented prior to the MA problem and the *F* key if it was different.

Nonwords only appeared once within the experiment, and they were presented in a different random order to each participant. For half of the nonword trials, the second word matched one of the three words presented before the MA problem. Further, the second nonword's position relative to the first three nonwords (i.e., whether the second nonword corresponded to the first, second, or third nonword presented before the problem) varied randomly across MA problems. Matches and mismatches between the second nonword and those seen before the MA problem varied equally across both horizontal and vertical presentations and across the working memory demands of the problems being performed.

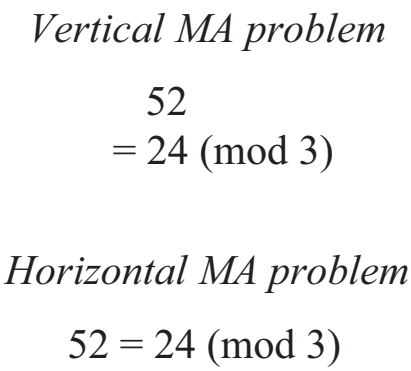

Figure 2. Example of vertical and horizontal modular arithmetic problem.

Results

MA. No MA problems were performed below 65% correct across all participants in the baseline condition. To preview, problem orientation did not result in significant differences in MA accuracy or in RTs across the single-task baseline and dual-task blocks. However, as we show below, this was not the case for secondary task performance. Because participants were instructed to allocate equal amounts of attention to both the math and the phonological task, decrements on either task as a function of problem orientation speaks to the differential working memory resources that horizontal and vertical MA problems use (Trbovich & LeFevre, 2003).

Accuracy and corresponding RTs for MA problems to which responses were correct were compared in a 2 (block: single-task baseline, dual-task) × 2 (problem working memory demand: low-

demand, high-demand) × 2 (problem orientation: horizontal, vertical) within-subjects design.

In terms of problem-solving accuracy, this analysis revealed a main effect of problem demand, $F(1, 23)=12.69$, $p < .01$, $\eta_p^2 =.36$, and a Block × Working Memory Demand interaction, $F(1, 23)=9.16$, $p < .01$, $\eta_p^2 =.29$. The low-demand problems did not differ in accuracy from the single-task ($M = 96.6\%$, $SE = 1.2\%$) to the dual-task block ($M = 98.2\%$, $SE = 0.7\%$). The high-demand problems were performed less accurately in the dual-task ($M = 85.9\%$, $SE = 3.0\%$) in comparison to the single-task block ($M = 91.9\%$, $SE = 2.3\%$; CI: 87.2%–96.6%; $d = .46$).

Analysis of RTs revealed a main effect of problem difficulty, $F(1, 23)=908.5$, $p < .01$, $\eta_p^2=.98$, in which the low-demand problems were performed faster than high-demand problems, and a Block × Difficulty interaction, $F(1, 23)=5.23$, $p < .04$, $\eta_p^2 =.19$, in which high-demand problem RTs increased from the single-task baseline to the dual-task block, whereas low-demand RTs decreased (Table 2). No other main effects or interactions reached significance (Fs ≤ 1).

Phonological secondary task. Accuracy and corresponding RTs for phonological secondary task problems to which responses were correct were analyzed as a function of the working memory demands and orientation of the MA problem they were paired with in a 2 (MA problem working memory demand: low-demand, high-demand) × 2 (MA problem orientation: horizontal, vertical) ANOVA. There were main effects of MA problem working memory demand, $F(1, 23)=15.33$, $p < .01$, $\eta_p^2 =.40$, and MA problem orientation, $F(1, 23)=9.55$, $p < .01$, $\eta_p^2 =.29$, which were qualified by a significant working memory demand by problem orientation interaction, $F(1, 23)=6.27$, $p < .02$, $\eta_p^2 =.21$.

There was no difference in phonological accuracy when performing low-demand ($M = 86.5\%$, $SE = 2.7\%$) or high-demand ($M = 86.5\%$, $SE = 2.1\%$) vertical MA problems. In contrast, when performing horizontal MA problems, phonological task accuracy was significantly higher for low-demand ($M = 85.9\%$, $SE = 2.4\%$) in comparison to high-demand problems ($M = 73.4\%$, $SE = 2.8\%$; CI: 67.5%–79.3%; $d = 0.96$). Phonological task accuracy while performing high-demand horizontal MA problems was also significantly worse than when performing low-demand and high-demand vertical MA problems ($d = 0.95$ and $d = 0.79$, respectively). Thus, phonological secondary task accuracy was lowest when individuals also performed high-demand horizontal MA problems, suggesting that horizontal high-demand problems and the phonological task are competing for the same verbal working memory resources (Baddeley, 1997).

In terms of phonological secondary task RTs, a 2 (MA problem working memory demand: low demand, high demand) × 2 (MA problem orientation: horizontal, vertical) ANOVA produced a main effect of difficulty, $F(1, 23)=5.14$, $p < .04$, $\eta_p^2 =.18$, in which individuals were slower to respond to the phonological secondary task when it was performed with a high-demand in comparison to a low-demand MA problem and a marginal Difficulty × Direction interaction, $F(1, 23)=3.02$, $p = .096$, $\eta_p^2 =.12$. Although not significant, this interaction parallels the phonological accuracy data in that the slowest phonological task RTs were seen in association with the performance of the horizontal high-demand MA problems (see Table 2).

Discussion

Adding a phonological memory load to MA execution led to performance decrements (primarily reflected in a decrease in secondary task accuracy) only when the MA problems being performed were high in working memory demands and presented in a horizontal orientation. Because participants were instructed to perform both the MA and the phonological secondary tasks equally well, errors in either task are evidence of disruption in working memory (Ashcraft & Kirk, 2001). This finding, similar to that reported by Trbovich and LeFevre (2003), suggests that high-demand horizontal (more so than vertical) MA problems and the phonological secondary task were competing for the same pool of verbal resources. More important, these findings establish the conditions to test whether stereotype-threat-induced failure is strongest for problems that rely most heavily on verbal working memory resources.

Experiment 3A

Women performed horizontally presented or vertically presented MA problems that were low or high in working memory demands in both a baseline and a stereotype threat condition. Prior to the baseline, all participants were told to perform their best. Prior to the stereotype threat condition, individuals were informed that the research was investigating why men are generally better

Table 2
Mean Response Times as a Function of Block, Problem Working Memory Demand, and Problem Orientation in Experiment 2

	Single task		Dual task			
	MA		MA		Phonological	
Demand and orientation	*M*	*SE*	*M*	*SE*	*M*	*SE*
Low demand						
Horizontal	2,287	135	2,066	105	1,252	73
Vertical	2,239	110	2,100	123	1,308	62
High demand						
Horizontal	7,876	550	7,944	653	1,416	69
Vertical	7,357	566	8,097	620	1,371	61

Note. Mean response times are reported in real-time metric (i.e., ms). MA = modular arithmetic.

than women at math. Following math task performance, participants reported any thoughts, feelings, and worries they experienced while performing under stereotype threat.

Method

Participants. Thirty-three women qualified for study participation by using the same criteria as Experiment 1. Eighteen individuals were randomly assigned to the vertical MA condition and 15 participants to the horizontal MA condition. Individuals in the vertical and horizontal conditions did not differ in terms of their perceptions of their math skill (vertical: $M = 7.44$, $SE = 0.29$; horizontal: $M = 7.00$, $SE = 0.31$), $F(1, 31) = 1.1$, $p = .30$, or the importance they assigned to this skill (vertical: $M = 7.33$, $SE = 0.23$; horizontal: $M = 7.13$, $SE = 0.34$; $F < 1$).

Procedure. Participants provided informed consent and were introduced to the vertical or horizontal MA task. Individuals first performed a practice block consisting of eight problems (four low demand, four high demand) presented in a different random order to each participant.

Following the practice block, all participants completed a baseline consisting of 20 MA problems (10 low demand, 10 high demand) presented in a different random order to each participant. Individuals were simply informed to perform their best during the baseline condition—solving the problems as quickly as possible without sacrificing accuracy. All participants were then presented with the stereotype threat manipulation used in Experiment 1 (see Appendix).

All individuals next completed another block of 20 MA problems (10 low demand, 10 high demand) presented in a different random order to each participant (stereotype threat block). Problems within the baseline and the stereotype threat block were counterbalanced across participants and were presented only once across the experiment. Thus, the presence of stereotype threat was manipulated within participants, with participants providing their own baselines.

Next, individuals completed a questionnaire intended to elicit their thoughts during the stereotype threat block of problems (Beilock et al., 2004). This questionnaire stated, "We all have several thoughts that run through our mind at any given time. Please describe everything that you remember thinking about as you performed the last set of modular arithmetic problems."

Participants completed two additional questionnaires designed to ensure that there were no differences (as a function of the orientation of the problems performed) in participants' perceptions of the importance of performing well on the MA problems or reported state anxiety following stereotype threat. Individuals completed the state form of the State–Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970), which consists of 20 questions that assess participants' feelings at a particular moment in time. Individuals responded to items (e.g., "I feel at ease") on scale ranging from 1 (*not at all*) to 4 (*very much so*). Following the STAI, participants responded to a question (on a 7-point scale) regarding their perceptions of the importance of performing at a high level on the last block of MA problems, ranging from 1 (*not at all important to me*) to 7 (*extremely important to me*). Individuals were then debriefed.

Results

Participants performing horizontal and vertical MA problems did not differ in their perceptions of the importance of performing well under stereotype threat. Participants in the vertical orientation ($M = 4.67$, $SE = 0.35$) and in the horizontal orientation ($M = 5.27$, $SE = 0.37$) condition reported that it was at least "moderately important" to perform well on these problems, $F(1, 31) = 1.36$, $p = .25$. Similarly, participants in the vertical orientation ($M = 33.22$, $SE = 1.6$) and in the horizontal orientation ($M = 37.00$, $SE = 2.7$) condition did not differ in reports of state anxiety, $F(1, 31)=1.53$, $p = .22$. Thus, any differences in MA performance under stereotype threat as a function of problem orientation reported below cannot be accounted for by differences in anxiety or perceived importance between the two problem orientation conditions.

Verbal Thought Questionnaire. Responses were divided into the following four categories:

1. Worries about the task or thoughts about confirming the stereotype threat manipulation (e.g., "I thought about how boys are usually better than girls at math so I was trying harder not to make mistakes [even though I did]" and "I was nervous in the last set because I found out that the study is to compare mathematical ability of guys and girls").

2. Thoughts regarding monitoring performance and its consequences (e.g., "I wanted to make sure I went as fast as possible but still get the answers right" and "I wish I was better at subtracting numbers in my head").

3. Thoughts related to carrying out the steps involved in performing the math problems (e.g., "I first saw whether the ones column could be subtracted without borrowing" and "I must subtract the two numbers and then see if the answer was divisible by the modular number given").

4. Thoughts unrelated to the experimental situation (e.g., "Walking home in the rain").

Two experimenters unaware of the hypotheses or experimental conditions independently coded the Verbal Thought Questionnaire data. Interjudge agreement was extremely high (97.8%), and thus one judge's coding was used for all responses.

On average, participants reported about three thoughts in total ($M = 2.82$, $SE = 0.24$). We next looked at the percentage of reported thoughts that fell into the categories outlined above. Because the thoughts questionnaire was open ended, we focused on proportion of thoughts (rather than total number of thoughts) reported to ensure that individual differences in the propensity to report thoughts in general were controlled across participants. However, the use of raw number of thoughts produced the same pattern of data as that reported below.

As seen in Table 3, with respect to the specific thought categories, 14.5% reflected worries or thoughts about confirming the stereotype threat manipulation, 34.9% were thoughts regarding monitoring their performance and its consequences, 32.4% were related to the steps involved in performing the math problems, and 18.3% were thoughts unrelated to the current experiment. Participants in the horizontal and

Table 3
Verbal Thoughts (by Type) Reported (in Percentages) for Experiments 3A (Stereotype Threat) and 3B (No Stereotype Threat)

Thought type	Experiment 3A	Experiment 3B
Worries about task/confirming stereotype	14.5	4.2
Monitoring performance and consequences	34.9	29.7
Steps in performing math problems	32.4	30.5
Unrelated to the experiment	18.3	35.5

vertical conditions did not significantly differ in the total number of thoughts reported ($F < 1$) or in the proportion of verbal reports across the categories (Category 1: $F < 1$; Category 2: $F(1, 31)=2.17$, $p = .15$; Categories 3 and 4: Fs < 1).

Thus, worries about the situation and monitoring performance and its consequences accounted for about half ($M = 49.4\%$, $SE = 6.5\%$) of participants' reported thoughts under stereotype threat. As seen below, the questionnaire data, when combined with MA performance, provide converging evidence that stereotype-induced consumption of working memory (especially phonological components of this system) is responsible for less-than-optimal performance in mathematical problem solving.

MA. As in the previous experiments, specific MA problems (and their corresponding RTs) that were not performed at least 65% correct across all participants in the baseline condition were removed from both the baseline and the stereotype threat blocks in the following analyses. Three problems and their corresponding RTs (out of the 80 total problems used in the baseline and stereotype threat blocks) were removed from the entire experiment.

Accuracy and corresponding (log-transformed) RT measures for problems to which responses were correct were compared in 2 (block: baseline, stereotype threat) $\times$ 2 (problem working memory demand: low demand, high demand) $\times$ 2 (problem orientation: horizontal, vertical) ANOVAs, with problem orientation as the between-subjects variable.

Analysis of accuracy revealed the anticipated three-way interaction, $F(1, 31)=4.12$, $p = .05$, $\eta_p^2 = .12$. As seen in Figure 3, the impact of stereotype threat was quite different depending on the working memory demand and the orientation of the problems being performed. For vertical problems, there was no Block $\times$ Problem Demand interaction ($F < 1$). In contrast, there was a significant Block $\times$ Problem Demand interaction for the horizontal problems, $F(1, 14)=7.70$, $p < .02$, $\eta_p^2 = .36$. Although the horizontal low-demand problems did not significantly differ in accuracy from the baseline ($M = 98.7\%$, $SE = 0.9\%$) to stereotype threat ($M = 100\%$, $SE = 0\%$) block, the horizontal high-demand problems were performed significantly less accurately in the stereotype threat block ($M = 81.2\%$, $SE =$

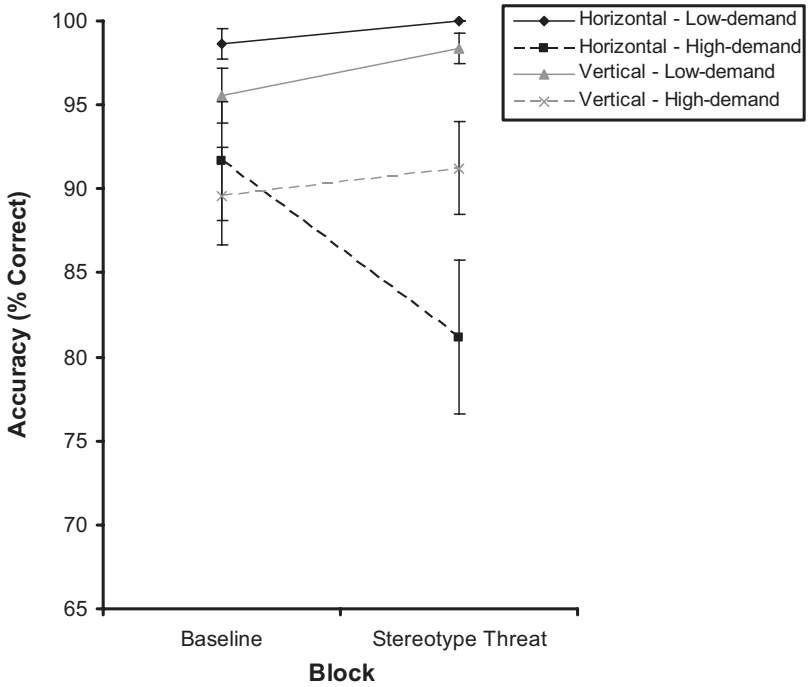

Figure 3. Accuracy (percentage correct) in the baseline and stereotype threat blocks for the low-demand and high-demand problems in the horizontal and vertical conditions in Experiment 3. Error bars represent standard errors.

4.6%) as compared with baseline ($M = 91.7\%$, $SE = 3.6\%$; CI: 84.0%–99.3% $d = 0.64$).

This pattern of data supports the prediction that stereotype threat targets the working memory resources on which horizontal high-demand problems rely for successful execution. Given that Experiment 2 and previous research (Trbovich & LeFevre, 2003) has shown that arithmetic problems presented in a horizontal format rely more on verbal resources than do vertically presented problems, this finding suggests that stereotype threat harms MA performance by co-opting the phonological resources that horizontal problems also use.

A three-way ANOVA on RTs also revealed a Block × Problem Working Memory Demand × Problem Orientation interaction, $F(1, 31)=9.68$, $p < .01$, $\eta_p^2 = .24$. A 2 (block: baseline, stereotype threat) × 2 (problem demand: low demand, high demand) ANOVA on vertical problem RTs revealed only a main effect of problem demand, $F(1, 17)=306.32$, $p < .01$, $\eta_p^2 = .95$, in which high-demand problem RTs were slower than low-demand problem RTs (see Table 4).

A similar ANOVA on horizontal problem RTs revealed a significant Block × Problem Demand interaction, $F(1, 14)=11.04$, $p < .01$, $\eta_p^2 = .44$. As seen in Table 4, while horizontal low-demand problem RTs decreased from the baseline to stereotype threat block, horizontal high-demand problem RTs increased, albeit not significantly.

Discussion

Women performed either horizontal or vertical MA problems that were low or high in working memory demands in both a baseline and a stereotype threat block. There were no differences as a function of block (i.e., baseline vs. stereotype threat) for vertical problem performance—regardless of problem working memory demand. However, this was not the case for the horizontal problems. Although the horizontal low-demand problems were not impacted by the introduction of a negative performance stereotype, the horizontal high-demand problems were performed significantly worse under stereotype threat in comparison to baseline conditions.

Individuals were also asked to report their thoughts during the stereotype threat block. Approximately half of these reported thoughts related to worries about the stereotype threat situation and to monitoring performance and its consequences. Unfortunately, off-line measures such as these cannot capture the intensity, duration, or precise timing of participants' thoughts. Thus, assessing direct relations between the number of thoughts reported on the verbal questionnaire and performance under stereotype threat is problematic. What the verbal reports do reveal is that participants did indeed report worries and performance concerns while under stereotype threat and, furthermore, that the prevalence of these thoughts did not differ as a function of problem orientation. This suggests that although all individuals experienced worries and verbal thoughts related to their performance under stereotype threat, these thoughts were only problematic for those individuals performing horizontally presented problems—problems that rely heavily on verbal working memory resources. Nonetheless, one might note that we have not demonstrated that individuals worry more under stereotype threat than in a no threat situation. To address this issue, we conducted a follow-up study. In Experiment 3B, women performed the exact same MA problems used in Experiment 3A in a no stereotype threat control condition and were asked to report the thoughts they had while performing the MA problems.

Experiment 3B

Method

Participants. Forty-two women qualified for study participation using the aforementioned criteria and were evenly split between horizontal and vertical problem groups.

Procedures. Individuals took part in the exact same design as those participants in Experiment 3A, with one exception. Prior to the second block of 20 MA problems (i.e., the posttest), individuals in Experiment 3B were not presented with the stereotype threat manipulation (see Appendix for the control information that participants in Experiment 3B received).

Table 4
Mean Response Times for Experiment 3A (Stereotype Threat) and Experiment 3B (No Stereotype Threat) as a Function of Block, Problem Working Memory Demand, and Problem Orientation

	Low demand				High demand			
	Baseline		Posttest		Baseline		Posttest	
Orientation	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Stereotype threat								
Horizontal MA	2,059	134	1,945	123	6,558	758	7,411	692
Vertical MA	2,192	127	2,199	118	7,930	549	7,212	469
No stereotype threat								
Horizontal MA	2,261	101	2,121	113	7,831	391	7,177	374
Vertical MA	2,133	121	1,980	107	7,315	358	7,600	450

Note. Mean response times are reported in real-time metric (i.e., ms). MA = modular arithmetic.

Results

We began by examining the perceptions of individuals performing the horizontal and vertical problems in Experiment 3B. Regardless of problem orientation, individuals did not differ in terms of their perceptions of the importance of performing well on the last block of problems, both reporting that it was at least "moderately important" to perform well on these problems (horizontal group: $M = 5.05$, $SE = 0.31$; vertical group: $M = 5.38$, $SE = 0.25$; $F < 1$). Similarly, horizontal ($M = 40.19$, $SE = 2.02$) and vertical problem ($M = 37.10$, $SE = 2.44$) participants did not differ in their reports of state anxiety ($F < 1$).

We next compared state anxiety and importance reports across Experiments 3A and 3B in a 2 (experiment: 3A stereotype threat, 3B control) × 2 (problem orientation: horizontal, vertical) design. In terms of state anxiety, there was no main effect of experiment, $F(1, 71)=2.47$, $p = .12$, or orientation, $F(1, 71)=2.34$, $p = .13$, and no Experiment × Orientation interaction ($F < 1$). In terms of importance, again there were no main effects of experiment or orientation (Fs < 1), and no Experiment × Orientation interaction, $F(1, 71)=2.14$, $p = .14$.

The lack of state anxiety differences under the stereotype threat and no threat conditions is consistent with the generally weak relationship between self-reported anxiety and impaired performance found in the stereotype threat literature (for a review, see Cadinu et al., 2005). Yet, at the same time, it counters the correlations between heightened levels of state anxiety and skill failure reported in the performance pressure and test anxiety literatures (Beilock et al., 2004; Tohill & Holyoak, 2000). These divergent patterns of results suggest differences between these types of failures. Nonetheless, nonverbal anxiety (measured through body posture, mannerisms, etc.) has been shown to increase under stereotype threat (Bosson, Haymovitz, & Pinel, 2004). Clearly, more work is needed to elucidate the precise role of general anxiety in stereotype threat.

Verbal Thought Questionnaire. Responses were divided by using the same four categories as Experiment 3A and were coded by the same experimenters used in Experiment 3A. Again, interjudge agreement was extremely high (98.8%), and thus the same judge's coding used in Experiment 3A was used for all responses.

On average, participants reported about four thoughts in total ($M = 3.90$, $SE = 0.34$), which was a somewhat larger total than that reported in Experiment 3A (i.e., $M = 2.82$), $t(73)=2.47$, $p < .05$, $d=.57$. As participants in Experiment 3A and 3B were taken from the same subject pool (albeit at different times), we are not sure why this difference occurred. However, it has been suggested that negative thoughts and worries are likely longer lasting, more intense, and less easy to dispel than are positive or neutral thoughts (Brosschot, & Thayer, 2003; Martin & Tesser, 1989, 1996). Given that our measure could not capture the intensity or duration of participants' thoughts, it is likely that the lower number of overall thoughts reported under stereotype threat reflects a preoccupation with a few intense thoughts (e.g., related to worries about the situation and its consequences) rather than more thoughts that were fleeting and less task related. Indeed, as can be seen in Table 3, a significantly larger proportion of individuals' reported thoughts in Experiment 3B as compared with Experiment 3A were unrelated to the task at hand, $t(73)=2.26$, $p < .03$, $d=.55$. In contrast, as will be seen below, the proportion of participants' worries and thoughts related to the performance situation and its consequences was significantly greater under stereotype threat than no threat conditions. And statistically controlling for such worries eliminated differences in math task performance under threat, supporting a causal role of verbal thoughts and worries in stereotype threat-induced failure.

Turning to the specific thought categories in Experiment 3B, as Table 3 reports, 4.2% of these reports reflected worries about the task, 29.7% were thoughts regarding monitoring their performance and its consequences, 30.5% were related to the steps involved in performing the math problems, and 35.5% were thoughts unrelated to the current experiment. As in Experiment 3A, participants in the horizontal and vertical conditions did not significantly differ in the total number of thoughts reported ($F < 1$) or in the proportion of verbal reports across the categories (Categories 1 and 2: Fs < 1; Category 3: $F(1, 40)=1.17$, $p = .29$; Category 4: $F < 1$).

Thus, worries about the task accounted for only 4% of participants' reported thoughts and together with monitoring performance and its consequences, these thoughts accounted for roughly one third of what was reported ($M = 34.0\%$, $SE = 5.0\%$). To explore how (and if) such reports differed from individuals in Experiment 3A performing the same MA problems under stereotype threat, we next compared the proportion of reported worries and thoughts about monitoring performance and its consequences of individuals in Experiment 3B to their stereotype threat counterparts in Experiment 3A in a 2 (experiment: 3A stereotype threat, 3B control) × 2 (problem orientation group: horizontal, vertical) design.

In terms of percentage of reported worries, this analysis revealed a main effect of experiment, $F(1, 71)=5.97$, $p < .02$, $\eta_p^2 = .08$. Individuals under stereotype threat (Experiment 3A) reported a significantly greater proportion of their thoughts being devoted to worrying than those under the no threat condition in Experiment 3B. There was neither a main effect of problem orientation nor an Orientation × Experiment interaction (Fs < 1). A similar pattern of results was seen for the proportion of thoughts regarding monitoring performance and its consequences, although the main effect of experiment ($F < 1$) as well as the main effect of problem orientation, $F(1, 71)=1.29$, $p = .26$, and their interaction, $F(1, 71)=1.44$, $p = .23$, was not significant. An Experiment × Problem Orientation ANOVA on the percentage of reported worries together with monitoring performance and its consequences also produced a main effect of experiment, $F(1, 71)=4.09$, $p < .05$, $\eta_p^2 = .06$. Again, the main effect of problem orientation, $F(1, 71) = 2.02$, $p = .16$, and the Problem Orientation × Experiment interaction, $F(1, 71)=1.25$, $p = .27$, was not significant. Thus, individuals performing MA problems under the no stereotype threat control condition of Experiment 3B devoted a significantly lower portion of their thoughts to worrying about the situation and monitoring performance and its consequences than those performing the same problems under stereotype threat in Experiment 3A.

MA. Four problems were performed below 65% correct across all participants in the baseline condition, and thus these problems were eliminated from the analyses. Next, we analyzed problem-solving accuracy and log-transformed RTs for problems answered correctly. A 2 (block: baseline, posttest) × 2 (problem demand: low demand, high demand) × 2 (problem orientation: horizontal, vertical) ANOVA on accuracy revealed only a main effect of problem demand, $F(1, 40)=31.21$, $p < .01$, $\eta_p^2 = .44$, in which the

low-demand problems (baseline— horizontal: $M = 94.8\%$, $SE = 1.5\%$; vertical: $M = 97.6\%$, $SE = 1.2\%$; posttest—horizontal: $M = 96.2\%$, $SE = 1.5\%$; vertical: $M = 96.2\%$, $SE = 1.3\%$) were performed more accurately than the high-demand problems (baseline—horizontal: $M = 90.5\%$, $SE = 2.1\%$; vertical: $M = 87.2\%$, $SE = 3.0\%$; posttest—horizontal: $M = 89.5\%$, $SE = 2.5\%$; vertical: $M = 84.7\%$, $SE = 2.7\%$).

A similar ANOVA on RTs produced main effects of problem demand, $F(1, 40)=1761.77$, $p < .01$, $\eta_p^2 = .98$, and block, $F(1, 40)=8.45$, $p < .01$, $\eta_p^2 = .17$, which were qualified by a Problem Demand × Block interaction, $F(1, 40)=4.60$, $p < .04$, $\eta_p^2 = .10$. As seen in Table 4, the low-demand problems were performed faster than the high-demand problems. However, this difference was greater in the posttest than in the baseline block.

Moreover, if one compares RTs on the types of problems shown to be impacted by stereotype threat (i.e., high-demand problems) across Experiment 3A and Experiment 3B in a 2 (block: baseline, posttest) × 2 (problem orientation: horizontal, vertical) × 2 (experiment: 3A stereotype threat, 3B control) ANOVA, a significant three-way interaction obtains, $F(1, 71)=12.52$, $p < .01$, $\eta_p^2 = .15$. For the vertical high-demand problem RTs, a 2 (block: baseline, posttest) × 2 (experiment: 3A, 3B) ANOVA revealed no main effect of block ($F = 1.2$) or experiment ($F < 1$) and no Block × Experiment interaction, $F(1, 37)=3.68$, $p = .06$. A similar analysis of horizontal high-demand problem RTs revealed a significant Experiment × Block interaction, $F(1, 34)=10.78$, $p < .01$, $\eta_p^2 = 0.24$. Although horizontal high-demand RTs decreased from the baseline to the posttest in Experiment 3B, these same RTs increased from the baseline to the stereotype threat block in Experiment 3A. However, the simple effects did not reach significance.

A 2 (block: baseline, posttest) × 2 (problem orientation: horizontal, vertical) × 2 (experiment: 3A stereotype threat, 3B control) ANOVA on high-demand problem accuracy also revealed a significant three-way interaction, $F(1, 71)=3.98$, $p < .05$, $\eta_p^2 = .05$. For the vertical high-demand problems, a 2 (block: baseline, posttest) × 2 (experiment: 3A, 3B) ANOVA revealed no main effects (Fs < 1). In contrast, the same ANOVA for the horizontal high-demand problems revealed a significant Block × Experiment interaction, $F(1, 34)=4.21$, $p < .05$, $\eta_p^2 = .11$. Although horizontal high-demand problem accuracy significantly decreased in Experiment 3A from the baseline block ($M = 91.7\%$, $SE = 3.6\%$) to the stereotype threat block ($M = 81.2\%$, $SE = 4.6\%$; $d = 0.64$), accuracy for the same problems in Experiment 3B did not (baseline: $M = 90.5\%$, $SE = 2.1\%$; posttest: $M = 89.5\%$, $SE = 2.5$).

In a final set of analyses we explored a more direct link between participants' reported worries and the specific type of performance shown in Experiment 3A (and Experiment 1) to be most impacted by stereotype threat: High-demand horizontal MA problem accuracy. Specifically, we performed the same 2 (block: baseline, posttest) × 2 (experiment: 3A, 3B) ANOVA on horizontal high-demand problem accuracy presented above and added as a covariate the proportion of reported worries together with monitoring performance and its consequences. To the extent that worries and thoughts about performance consequences underlie stereotype threat, covarying out these thoughts should render the significant Block × Experiment interaction reported above nonsignificant. This is exactly what was found, $F(1, 33)=2.8$, $p = .10$.

Discussion

Under the no stereotype threat conditions of Experiment 3B, women performed at a high level on the MA tasks, regardless of problem orientation or demand. Moreover, in comparison to women performing the same problems under stereotype threat in Experiment 3A, a significantly lower proportion of individuals' reported thoughts in Experiment 3B were related to worries and thoughts about the situation and its consequences. Finally, the critical interaction of experiment and problem block for the type of performance shown to be most strongly impacted by stereotype threat across the first several studies in the current work (i.e., horizontal high-demand MA accuracy) was rendered nonsignificant when worries and thoughts about performance and its consequences was taken into account.

Taken together, the first three experiments suggest stereotype threat causes individuals to worry about their performance and its consequences and harms those math problems most reliant on verbal working memory resources. These studies provide the most comprehensive account to date of the specific mechanisms by which stereotype threat has its impact. In Experiment 4, we use this knowledge to engineer conditions in which those problems most impacted by stereotype threat (i.e., horizontally presented MA problems) should be unaffected by the introduction of a negative performance stereotype.

Experiment 4

Women were trained on 636 horizontal MA problems and then exposed to stereotype threat (i.e., the same negative stereotype regarding women and math used in the previous experiments). Problems within the training session occurred 48 times each (multiple repeats) or only once (no repeats), and were either low or high in working memory demands. To the extent that horizontal problems fail because stereotype threat co-opts the resources on which such problems rely, then extensively practicing these problems to the point where they are not heavily dependent on such resources should alleviate the negative impact of stereotype threat.

According to Logan's (1988) instance-based theory of how mental arithmetic is learned, a rule-based algorithm is initially used to solve unpracticed MA problems. That is, novel problem solutions are dependent on the explicit application of a capacity-demanding process that must be maintained and controlled on-line in working memory during execution. With practice on particular problems, the reliance on this procedure decreases and past instances of problem solutions are retrieved directly or "automatically" from long term memory into working memory (similar to how one's multiplication tables might be retrieved from memory), whereas new problems continue to rely on the algorithm. If stereotype threat reduces the working memory capacity needed to correctly solve horizontal problems, then regardless of how many different problems individuals have been exposed to, only problems that have been practiced enough to produce instance-based answer retrieval should be inoculated against stereotype threat—a minimum of 36 exposures according to Klapp, Boches, Trabert, and Logan (1991). New horizontal problems that have not been repeatedly practiced should continue to rely on algorithmic computation and the maintenance of intermediate problem steps as phonological codes. These new problems should be harmed by

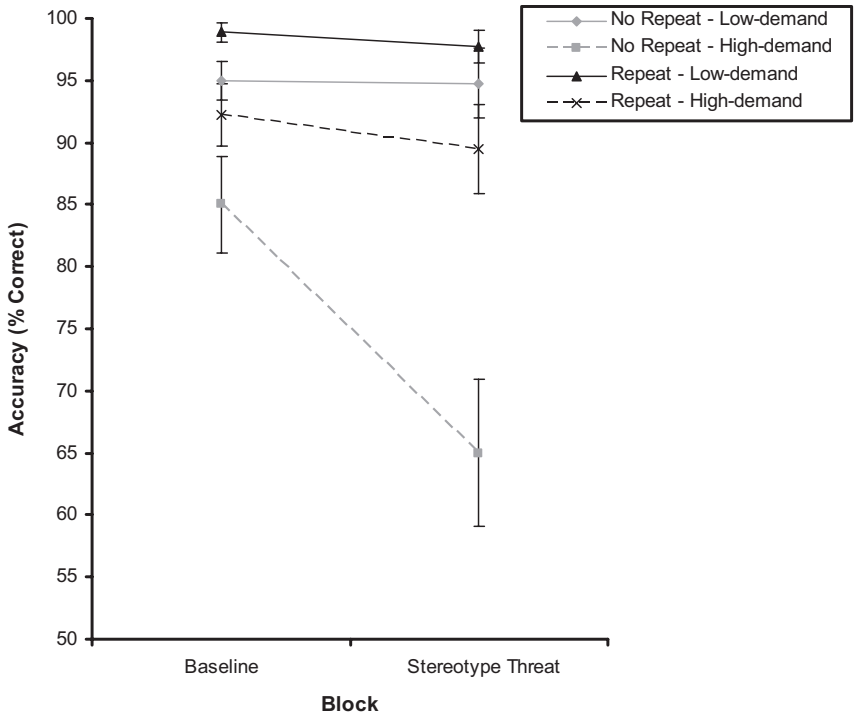

Figure 4. Accuracy (percentage correct) in the baseline and stereotype threat blocks for the multiple repeat and no repeat horizontal problems in Experiment 4. Error bars represent standard errors.

stereotype threat provided they are working memory dependent enough to be impacted when such resources are consumed.

Method

Participants. Thirty women meeting the same criteria used in the above experiments participated.

Procedure. Participants provided informed consent and were introduced to MA. All problems were presented in a horizontal orientation. Individuals first performed an eight problem practice block (four low demand, four high demand) presented in a different random order to each participant.

Following the practice block, participants completed three training blocks of 212 problems each. Within each block, 12 problems (6 low demand, 6 high demand) were presented 16 times each (multiple repeats) and 20 problems (10 low demand, 10 high demand) were presented once (no repeats). Problems were presented in a different random order to each participant. Thus, across the three training blocks, 12 multiple repeat problems were presented 48 times each and 60 no repeat problems were presented once.

Participants then completed two blocks of 24 problems each. The first block (baseline) appeared to be another series of training problems. Prior to the second block of problems (stereotype threat), all participants were given the same negative stereotype used above. Thus, stereotype threat was manipulated within participants with participants providing their own baselines. Both the baseline and stereotype threat blocks consisted of the 12 problems (6 low demand, 6 high demand) presented 48 times each during training and 12 problems (6 low demand, 6 high demand) not previously presented. Problems were presented in a different random order and counterbalanced across participants. Following the completion of the stereotype threat block, participants were completely debriefed.

Results

As in the previous experiments, specific MA problems (and their corresponding log-transformed RTs) that were not performed at least 65% correct across all participants in the baseline condition were removed from both the baseline and stereotype threat blocks in the following analyses. Seven problems and their corresponding RTs (out of the 48 total problems used in the baseline and stereotype threat blocks) were removed.

Next, accuracy and RTs for correct problems were analyzed in separate ANOVAs with a 2 (block: baseline, stereotype threat) $\times$ 2 (problem repetition: no repeat problems, multiple repeat problems) $\times$ 2 (problem working memory demand: low demand, high demand) design. As seen in Figure 4, in terms of accuracy, a significant Block $\times$ Problem Repetition $\times$ Problem Working Memory demand interaction obtained, $F(1, 29)=6.13$, $p < .02$, $\eta_p^2=.17$.

This three-way interaction was examined by analyzing the heavily practiced (multiple-repeat) problems and novel (no repeat) problems separately. A 2 (block: baseline, stereotype threat) $\times$ 2 (problem demand: low demand, high demand) ANOVA on the multiple repeat problems revealed no Block $\times$ Problem Demand interaction ($F < 1$). The same ANOVA on the no repeat problems revealed a significant Block $\times$ Problem Demand interaction, $F(1, 29)=11.11$, $p < .01$, $\eta_p^2 = .28$. Accuracy for the no repeat low-demand problems did not differ between the baseline ($M = 95.0\%$, $SE = 1.5\%$) and stereotype threat block ($M = 94.8\%$, $SE =$

Table 5
Mean Response Times for Experiment 4 as a Function of Block, Problem Working Memory Demand, and Problem Type

	Low demand				High demand			
	Baseline		Posttest		Baseline		Posttest	
Problem type	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
No repeat	1,872	148	1,826	96	6,666	679	7,751	804
Multiple repeat	1,115	50	1,156	63	2,132	183	2,163	144

Note. Mean response times are reported in real-time metric (i.e., ms).

2.8%). Accuracy for the no repeat high-demand problems significantly declined from the baseline ($M = 85.0\%$, $SE = 3.9\%$) to the stereotype threat block ($M = 65.0\%$, $SE = 5.9\%$; CI: 52.8%–77.2%; $d = 0.70$).

The analysis of RT data did not alter the conclusions supported by the accuracy analysis.[2] A 2 (block: baseline, stereotype threat) × 2 (problem repetition: no repeat problems, multiple repeat problems) × 2 (problem working memory demand: low-demand, high-demand) ANOVA on RTs revealed main effects of problem repetition, $F(1, 26) = 139.94$, $p < .01$, $\eta_p^2 = .84$, and problem demand, $F(1, 26) = 144.14$, $p < .01$, $\eta_p^2 = .85$, which were qualified by a significant Repetition × Demand interaction, $F(1, 26) = 56.97$, $p < .01$, $\eta_p^2 = .69$. No Block × Repetition × Problem Demand interaction was observed, $F < 1$.

As seen in Table 5, problem demand level had more of an effect on RTs for no repeat problems than for the multiple repeat problems. This indicates that practicing the multiple repeat problems reduced the initial time differences (as a function of problem demand) in their solution. Such a finding is consistent with the shift from algorithmic execution to direct answer retrieval from memory proposed by Logan's (1988) theory of instance-based automaticity. Once problems are repeatedly practiced to the point that their answers are retrieved directly from long-term memory, the working memory demands of the initial algorithmic computations should not markedly impact RTs because the algorithm is no longer being computed on-line as a means to derive the answer. Nonetheless, repeated high-demand problems did yield longer RTs than did repeated low-demand problems, suggesting that there was at least some degree of nonautomatic answer retrieval. However, the relatively fast RTs for the repeat problems (compared with the no repeat problems), coupled with the Repetition × Demand interaction, suggests that a majority of the repeated problems were answered via direct answer retrieval, which inoculated them against stereotype threat. Finally, the lack of a three-way interaction in RTs suggests the accuracy results are not the product of speed–accuracy trade-off. In fact, as accuracy for the high-demand no repeat problems declined from the baseline to the stereotype threat block, RTs increased, although not significantly.

Discussion

Performance of horizontally presented MA problems practiced 48 times each (multiple repeats), and thus not heavily reliant on working memory, did not fail under stereotype threat. Problems presented only once (no repeats) did. Furthermore, these failures were limited to the no repeat problems that placed the heaviest demands on verbal working memory.

Recent work has demonstrated that making individuals aware of performance stereotypes and their consequences can limit the occurrence of stereotype threat (Johns, Schmader, & Martens, 2005). Although this may be useful in teaching settings, it may be hard to dissuade someone of a robust and persistent performance stereotype in a threatening testing situation. The current experiment provides another route to such ends by leveraging knowledge about the causal mechanisms by which stereotype threat impacts performance to devise a training regimen to alleviate performance decrements. Thus, the current findings reaffirm the adage that "practice makes perfect," and further, they suggest an addendum to this statement. Practice not only makes perfect, but practice (provided that it creates less reliance on working memory) makes skills robust to stereotype threat effects. One might wonder whether repeatedly practicing problems is really a form of effective practice given that the specific problems individuals perform on high-stakes tests are often unknown ahead of time. However, the majority of math problems on such tests involve basic algebraic facts and mathematical procedures. Further, careless mistakes on these types of basic operations likely contribute to less-than-optimal performance in a variety of testing situations (Ashcraft & Kirk, 2001; Beilock & Carr, 2005; Beilock et al., 2004). Thus, practice designed to alleviate the working memory demands of the subcomponents of the problems one encounters should be an efficacious training strategy.

In Experiment 4, we demonstrated how an understanding of the processes by which stereotype threat operates can be used to attenuate unwanted performance decrements in math. In our final experiment, we take the opposite tactic: We use our understanding of the mechanisms underlying stereotype threat effects in cognitively demanding tasks such as math problem solving to predict when unwanted performance decrements may spill over onto unrelated tasks.

Experiment 5

Women performed horizontal MA problems under stereotype threat followed by either a verbal or a spatial computerized two-

[2] Three participants were not included in the RT analyses for Experiment 4 because they provided no correct responses for the no repeat high-demand problems under stereotype threat. Thus, there were no RTs (associated with correct problem responses) to include in these analyses.

back working memory task (Gray, 2001). If stereotype threat most strongly impacts verbal processing resources, and this impact does not immediately subside when performance on the stereotype-threat-related task is finished, then individuals should perform more poorly on a verbal, in comparison to a spatial, two-back task following stereotype threat even though this task is unrelated to the negative performance stereotype. Indeed, stereotype threat may "spill over" onto subsequent tasks that use the same processing resources despite the fact that such tasks are not implicated by the negative stereotype.

The above hypothesis is not anticipated by existent stereotype threat work, which assumes that threat effects are confined to the domain implicated by the stereotype in question (Steele, 1997). Nonetheless, if stereotype threat exerts its impact in tasks such as math problem solving by consuming working memory, and especially verbal resources, the prediction that stereotype threat might spill over onto tasks unrelated to the performance stereotype (yet dependent on the same cognitive resources impacted by stereotype threat) seems plausible. Moreover, demonstrations in the resource depletion literature that effortful expenditures of self-regulation can impact later tasks in different domains (e.g., Muraven & Baumeister, 2000; Richeson & Trawalter, 2005; Schmeichel, Vohs, & Baumeister, 2003), give such a prediction more credence.

It should be noted that the resource depletion literature focuses on the depletion of a general pool of cognitive resources following self-regulation (e.g., evoking the metaphor of a muscle that fatigues with use; Muraven & Baumeister, 2000). In the current work, we postulate that the consumption of a particular type of asset (i.e., verbal processing resources) by stereotype threat will impact other tasks dependent on this resource as well. Thus, we use our understanding of how stereotype threat affects performance to make particular predictions about the types of subsequent tasks that might be impaired. Toward this end, we stray away from the resource depletion literature's assertions regarding an undifferentiated nature of depletion (Muraven & Baumeister, 2000) and instead predict focused consequences of stereotype threat. Thus, in addition to testing a novel and unanticipated prediction in the stereotype threat literature concerning spillover, our final experiment puts forward a more specific mechanism of failure than what is hypothesized in the resource depletion literature to date. As such, this work serves to forage new ground in skill failure under stereotype threat and sheds light on possible mechanistic routes to depletion more generally.

Working Memory Pilot Test

Gray (2001) has previously reported that the two-back tasks used in Experiment 5 are well matched for difficulty. Nonetheless, we felt it was important to demonstrate this in our sample as well. Thus, we began by establishing (via pilot test in a no stereotype threat, control situation) that these tasks were relatively well equated in terms of performance under no stereotype threat control conditions.

Method

Participants. Twenty-seven women qualified for participation by demonstrating adequate performance on the two-back tasks (i.e., at least 70% accuracy). Fifteen women performed the verbal two-back task and 12 performed the spatial two-back task.

Two-back tasks. The verbal and spatial two-back tasks were implemented by using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993) on a Macintosh Quadra. In all cases, the presentation of the stimuli was the same, but the nature of the task (spatial vs. verbal) was manipulated by instructions presented to participants on the computer and read aloud by the experimenter. Individuals were told to indicate whether a stimulus item presented on the current trial matched the item presented two trials previously by using either the *S* key (same stimuli) or the *D* key (different stimuli). The stimuli were comprised of a cluster of identical letters (e.g., *b*s, *k*s) inside a 5.4-cm square presented in one of six different spatial locations in an ellipse around the center of the monitor against a background of random letters. On each trial, the target was presented (500 ms), followed by a 2,500-ms period during which only the background appeared. Thus, participants had 3 s to indicate their response before the next trial (a tone indicated participants' failure to respond and the trial was scored as an error).

Participants in the verbal condition determined whether the letters of the current stimulus trial matched the letters presented two trials earlier (ignoring the physical location of those presentations), whereas those in the spatial condition indicated whether the presentation location of the current stimulus trial matched the same location as the stimuli presented two trials earlier (ignoring the letters presented in those presentations). The first response trial occurred following the third stimulus presentation (in which the stimulus was compared with the stimulus presented in the first stimulus presentation). Participants were given an initial practice session of 10 response trials to ensure that they understood the task (repeated if necessary). Later, they completed the critical session, consisting of 100 response trials. In each session (practice and critical), 30% of the trials were "same" trials, and the remaining 70% were "different" trials.

Results

Accuracy and RT measures for correct critical trials were analyzed. There were no differences in either two-back accuracy (verbal: $M = 86.6\%$, $SE = 1.9\%$; spatial: $M = 84.3\%$, $SE = 3.0\%$) or RTs (verbal: $M = 830$ ms, $SE = 42$ ms; spatial: $M = 861$ ms, $SE = 60$ ms), as a function of which two-back task individuals performed (Fs < 1). We now turn to the main section of Experiment 5 in order to explore whether performing MA problems under stereotype threat prior to two-back performance altered this above pattern of results.

Primary Experiment

Method

Participants. Thirty-three women qualified for study participation with the same criteria as the pilot test. Fifteen women performed the MA task followed by the verbal two-back task. Eighteen women performed the MA task followed by the spatial two-back task. Two-back task version was randomly assigned. An additional 3 participants qualified but were not retained as study participants because they failed to spend an adequate amount of time reading the stereotype threat manipulation (i.e., <30 s). In contrast to previous studies in the current work, the stereotype

threat manipulation in Experiment 5 was presented following general task instructions. Thus, it was possible that individuals who believed they understood the general task instructions would fail to spend sufficient time reading the stereotype threat manipulation. This was not likely in the previous studies because the stereotype threat manipulation was separated from the general task instructions, occurring after an initial baseline block of MA problems.

Procedure. Participants provided informed consent and, as in the pilot test, were introduced to and completed the 10-trial two-back practice task, either spatial or verbal (on the basis of condition assignment). They completed this practice first in order to ensure that they understood the two-back task prior to completing subsequent tasks. Next, participants were moved to a second computer and were introduced to MA. All individuals were then given the stereotype threat scenario (see Appendix) and subsequently performed 20 high-demand, horizontal problems (i.e., the problems demonstrated previously to show stereotype threat effects).

Following MA problem completion, individuals were moved back to the computer on which they had practiced the two-back task. They then performed the critical 100 trials of the same version of the two-back task they had practiced prior to the MA problems. Afterwards, they were thanked and debriefed.

Results

MA. Individuals performed only horizontal, high-demand problems (accuracy: $M = 85.3\%$, $SE = 2.6\%$; RT: $M = 6345$ ms, $SE = 469$ ms). To ensure that our stereotype threat manipulation was successful, we looked next to the performance measure that had consistently demonstrated stereotype threat effects in the above experiments—horizontal high-demand problem accuracy.

We began by comparing horizontal high-demand accuracy in Experiment 5 with accuracy in the same type of problems in the other experiments in which stereotype threat was manipulated at low levels of practice (i.e., the stereotype threat group in Experiment 1 and the stereotype threat block in Experiment 3). This analysis revealed no difference in horizontal high-demand problem accuracy as a function of experiment ($F < 1$). To confirm that this lack of accuracy difference reflected equally low levels of performance under stereotype threat, we next compared performance on these horizontal high-demand problems under stereotype threat with performance on the same type of problems under no threat conditions (i.e., posttest performance for the control group in Experiment 1 and Experiment 3B). A significant main effect of stereotype threat was found, $F(1, 96) = 8.56$, $p < .01$, $\eta_p^2 = .08$. Across experiments, performance on the horizontal high-demand problems under stereotype threat ($M = 83.2\%$, $SE = 2.0\%$) was significantly lower than performance on the same problems under no threat conditions ($M = 91.5\%$, $SE = 1.5\%$).

Two-back task. As in the pilot, accuracy and RT measures for correct trials were analyzed, revealing faster and more accurate spatial than verbal two-back task performance. The RT difference was significant (verbal: $M = 1{,}087$ ms, $SE = 59$ ms; spatial: $M = 895$ ms, $SE = 49$ ms), $F(1, 31) = 6.33$, $p < .02$, $\eta_p^2 = .17$. The accuracy difference was not (verbal: $M = 87.3\%$, $SE = 1.7\%$; spatial: $M = 89.0\%$, $SE = 1.5\%$; $F < 1$). Because individuals attempted to perform the task as quickly and accurately as possible, either accuracy or RT measures can be used as an index of performance.

How did two-back performance following stereotype threat in MA compare with two-back performance under control conditions? To address this, we examined two-back RT and accuracy in a 2 (task: verbal, spatial) × 2 (experiment: control pilot, stereotype threat) design. As seen in Figure 5, a Task × Stereotype Threat interaction obtained for RT, $F(1, 56) = 4.38$, $p < .05$, $\eta_p^2 = .07$. The verbal two-back task was performed significantly slower than the spatial two-back task following stereotype threat in MA. This did not occur when the two-back task was not preceded by stereotype threat performance. There were no significant effects for the accuracy analyses.

If stereotype threat spills over onto subsequent tasks, then those who performed the poorest under stereotype threat in math should also show the poorest performance on the two-back task. And, if stereotype threat exerts its impact by drawing most heavily on phonological resources, then the relation between MA performance under stereotype threat and the two-back task should hold most strongly for the verbal two-back task. Accordingly, we examined the relation between MA and two-back task performance as a function of the type of two-back task individuals performed. In this context, better performance can be revealed by greater accuracy, faster RTs, or both. Thus, we conducted three sets of multiple regressions in which two-back task performance (where spillover was exhibited) was regressed on MA performance, the type of two-back task (dummy coded), and their interaction (the key prediction). The three regression analyses examined performance on MA and the two-back task by using standardized accu-

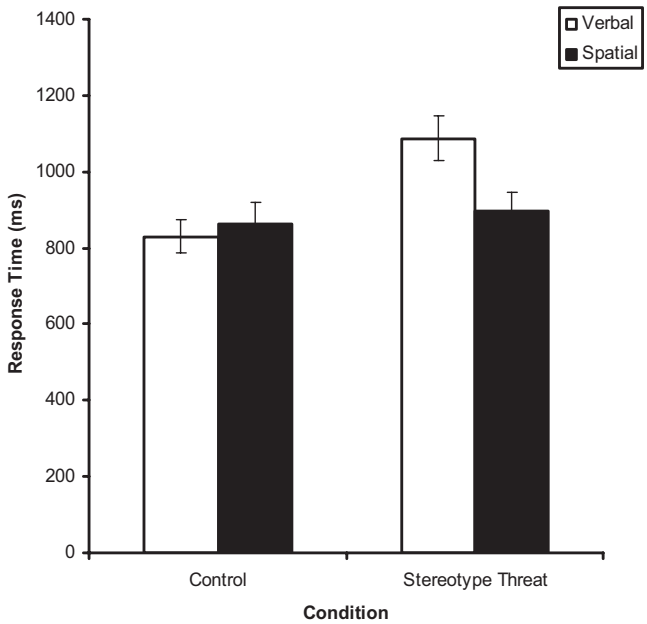

Figure 5. Verbal and spatial two-back task reaction time (ms) in the pilot (control) and stereotype threat conditions in Experiment 5. Error bars represent standard errors.

racy, standardized RTs, and a composite of the two (standardized accuracy minus standardized RTs).[3]

With RT as an index of performance, there was a main effect of two-back task type ($\beta = -.36$) $t(29) = 2.34$, $p < .03$, a marginal main effect of MA RT ($\beta = .28$) $t(29) = 1.84$, $p < .08$, and the predicted interaction between the two ($\beta = -.32$) $t(29) = 2.13$, $p < .05$. As expected, the relation between RTs for MA and the two-back task was significant for those completing the verbal two-back task ($r = .64$, $p < .01$) but not for those completing the spatial two-back task ($r = -.05$, *ns*). When using accuracy as an index of performance, there was a main effect of MA performance ($\beta = .44$) $t(29) = 2.27$, $p < .04$, and a marginal interaction of MA performance and two-back task type ($\beta = -.31$) $t(29) = 1.61$, $p = .12$. Although not reliable at conventional levels, this latter outcome reflects that the relation between MA accuracy and two-back task accuracy was, as expected, significant for the verbal two-back task ($r = .57$, $p < .03$) but not for the spatial two-back task ($r = .15$, *ns*).

Finally, we conducted a multiple regression analysis by using composite measures that captured both accuracy and RTs. In many ways, the composite reflects the best index of performance because it simultaneously takes into account both accuracy and latency. In this analysis, there was a main effect of two-back task ($\beta = .35$) $t(29) = 2.30$, $p < .03$, which was qualified by its predicted interaction with MA performance ($\beta = -.44$) $t(29) = 2.91$, $p < .01$. As found above for both accuracy and RT separately, the relation (now with the composite approach) between performance for MA and the two-back task was significant for the verbal two-back task ($r = .65$, $p < .01$) but not for the spatial two-back task ($r = -.26$, *ns*). Thus, regardless of whether performance was defined as accuracy, latency, or a composite of the two, those who performed worse on the MA task under stereotype threat performed more poorly on the subsequent two-back task—however, this relation only held for verbal two-back task performance.

Discussion

To our knowledge, Experiment 5 is the first demonstration that following underperformance on a stereotype-relevant task, subsequent task performance in a different domain is also negatively impacted—as long as the subsequent task depends heavily on the same type of working memory resources that stereotype threat also consumes. This stereotype threat spillover occurred despite the subsequent task being unrelated to the stereotype in question. That is, a math-related stereotype should not apply to verbal task performance. If anything, women might anticipate doing better in a verbal domain (e.g., Seibt & Forster, 2004). In summary, performance decrements were observed in a task performed subsequent to the stereotyped task, demonstrating how stereotype threat can spill over onto other activities not implicated by the stereotype in question.

General Discussion

Although stereotype threat has been demonstrated for many social groups and task types, its underlying causal mechanisms have received far less attention. The current work examined how negative performance stereotypes impact the cognitive resources necessary to successfully execute working memory intensive tasks such as mathematical problem solving. Results revealed that stereotype threat exerts its impact by co-opting working memory resources—especially phonological aspects of this system—needed for the successful performance of some types of math problems (e.g., horizontal high-demand problems) more than others (e.g., vertical low-demand problems). However, stereotype threat effects in the former problem type were alleviated by rendering the use of a working-memory-demanding computational algorithm unnecessary by repeated problem practice. Finally, we demonstrated that stereotype threat may not only impact performance in the domain implicated by the stereotype, but it can spill over onto subsequent, unrelated tasks that depend on the same processing resource that stereotype threat consumes. These novel findings not only provide insights into the cognitive underpinnings of stereotype threat but also reveal new circumstances when its effects are attenuated and propagated. Such knowledge contributes to our theoretical understanding of stereotype threat and speaks to how environmental factors (e.g., highlighting social group membership) can influence the working memory system. This is an issue that has not yet received adequate attention in the working memory literature (Miyake & Shah, 1999a) but is of import for researchers interested in developing models of working memory that capture the complexity of real-world performance.

Our work also provides evidence that stereotype threat induces task-related thoughts and worries (for converging evidence, see Cadinu et al., 2005) that target phonological aspects of working memory. Using Baddeley's (1986; Baddeley & Logie, 1999) multicomponent model as a framework, one could unpack verbal working memory into a phonological store capable of holding speech-based information and an articulatory control process based on inner speech mechanisms. It has been suggested that the temporary maintenance of intermediate steps, as well as the on-line updating of such information, may be especially dependent on such phonological resources (DeStefano & LeFevre, 2004). Thus, horizontal high-demand problems that prompt the intermediate steps of a borrow operation to be maintained as phonological codes and that require the updating of such information via articulatory control processes, may be particularly susceptible to stereotype threat-related worries—especially if such thoughts automatically capture phonological resources as part of their initial registration process.

It is, however, worth pointing out that there are a few studies that have found stereotype threat effects in tasks with spatial components. Martens, Johns, Greenberg, and Schimel (2006) recently demonstrated that when made aware of gender differences in spatial rotation ability, women performed worse than men on a test involving discriminating between figures in different spatial rotations. Also, Gonzales, Blanton, and Williams (2002) found stereotype threat effects on a task of math and spatial ability. However, though the above mentioned tasks do depend somewhat on spatial resources for successful execution, they also likely draw

[3] A difference score is used for the composite because better performance reflects greater accuracy and faster responses (hence, RTs are subtracted from accuracy). Also, because previous work that used the two-back tasks did not log transform the RT measures (see Gray, 2001), we did not include this transformation in our analyses. However, transforming the data would not have changed the pattern of results reported.

heavily on central executive and even verbal resources. The Gonzales et al. task, for example, involved general mathematical computations and the Martens et al. task involved discriminating between several different answer options at once, which likely taxed more than just spatial processing resources. Regardless of the specific subcomponents on which such tasks rely, the current work's demonstration of a heavy involvement of verbal resources in stereotype threat impairment does not exclude other subcomponents of the working memory system from being implicated in stereotype threat related failure. Rather, stereotype threat likely affects a combination of phonological loop functioning (via verbal thoughts and worries) and probably some central executive functioning (via attempts to suppress such thoughts and to focus on the task at hand). This leaves open the possibility that tasks with spatial components, but that also tax central executive or phonological resources, may show signs of stereotype threat as well—although we would argue that such failures should not be as pronounced as in tasks that depend more so on phonological aspects of the working memory system.

Is it possible that the current results could be accounted for solely by stereotype threat's impact on general executive control resources? There are a number of reasons why this notion seems highly unlikely. First, previous research (e.g., Trbovich & LeFevre, 2003) has shown that although horizontally presented problems are impacted most heavily by a phonological load, vertical problems are impacted more heavily by a spatial load, suggesting that it is horizontal problems' stronger reliance on phonological (rather than executive) resources that makes them susceptible to stereotype threat effects. Second, the vertical and horizontal problems presented in the current work were exactly the same—only orientation differed. And indeed, there was no difference in horizontal and vertical problem performance under single-task baseline conditions. Thus, it seems unlikely that one problem type would place heavier demands on central executive resources than another. Moreover, both types of high-demand problems (horizontal and vertical) involved carry operations that have been shown to implicate central executive resources. Thus, to the extent that stereotype threat or the phonological task used in Experiment 2 solely taxed executive resources, then both types of problems should have failed. Finally, not only did the verbal (but not spatial) two-back task in Experiment 5 show signs of stereotype threat induced spillover, the verbal two-back task was the only task that correlated with MA performance under stereotype threat. If general resource consumption could solely explain stereotype threat effects and their spillover, then a correlation between MA performance under stereotype threat and spatial two-back performance should exist, but there was not. In summary, an explanation for the current work's stereotype threat effects based exclusively on the taxing of general executive control resources does not seem tenable.

Because the purpose of this work was to examine how stereotype threat impacts performance rather than to generalize stereotype threat effects across different groups, only women were examined under stereotype threat. "Women and math" research is prevalent in the stereotype threat literature and has been identified as a priority in education (Kegel-Flom & Didion, 1995; Steele et al., 2002). Moreover, women comprise one of the largest stigmatized groups. Thus, understanding what leads to the underperformance of such a large segment of the population is very important. Nonetheless, the mechanisms of failure in the current work should extend to anyone who falls prey to stereotype threat effects, regardless of how these effects arise: African Americans (Steele & Aronson, 1995), Latinos (Gonzales et al., 2002), and even White men who are compared with Asians (Aronson et al., 1999).

Further, in the above experiments, the experimental blocks (i.e., the dual-task or stereotype threat blocks) always followed the baseline condition. This set order was necessary because it is impossible to obtain a baseline measure of one's math performance after being exposed to a negative performance stereotype. And, in order to keep the experiments as closely aligned as possible, we implemented this order in the other experiments as well. The possibility that the results reported above are due to order effects seems highly unlikely however because all problem types were exposed to the same order effects, yet only specific types of problems (e.g., high-demand horizontal problems) were adversely impacted by the dual-task and stereotype threat situations. That is, it would be hard to explain why some types of problems were more susceptible to failure than others as a function of order rather than as a function of the cognitive representation of the problems themselves. Moreover, having the baseline condition always precede the experimental block should only make it more difficult to find skill decrements. Individuals always had more practice by the time they reached each experimental block than they had at the time of the comparison baseline. Thus, it should be especially difficult to obtain performance decrements (relative to a less practiced baseline) at this point. Finally, in several of the studies presented above, individuals performing under control conditions showed dramatically different patterns of results than those performing under stereotype threat, which further weakens the likelihood that order effects played a role in the current work.

Stereotype Threat, Verbal Working Memory, and Withdrawal-Motivated States

The idea that stereotype threat most strongly impacts problems that rely heavily on verbal working memory may be related to work examining the impact of induced affective states on cognitive control. Specifically, it has been demonstrated that unpleasant (withdrawal-motivated) affective states impair verbal working memory yet improve spatial working memory (Gray, 2001; Gray, Braver, & Raichle, 2002). Withdrawal states have been shown to lead to greater right hemisphere activation in comparison to pleasant affective states (approach-motivated), which increase left hemisphere activation (specifically the prefrontal cortex and the amygdala; Davidson, Jackson, & Kalin, 2000). In addition, neuroimaging studies have shown that the active maintenance of verbal information depends more on the left prefrontal cortex, whereas the active maintenance of spatial information depends more on the right prefrontal cortex (Smith & Jonides, 1999). Thus, it may be that stereotype threat induces a negative affective state that not only leads to verbal thoughts and worries but also reduces the verbal working memory capacity available for any verbal information, whether necessary task information or situational worries (Beilock & Carr, 2005).

Stereotype Threat, Performance Under Pressure, and Math Anxiety

We began the current work by turning to the performance pressure and anxiety literatures for clues concerning how stereotype threat might compromise working memory in tasks such as math problem solving. What then are the relations between stereotype threat, choking under pressure, and math anxiety in this domain? There are at least two different components one can focus on in thinking about the similarities and differences between stereotype threat, choking under pressure, and math anxiety. The first concerns the environmental triggers that induce failure. The second concerns an understanding of how these failure mechanisms play out in the cognitive control systems that support performance in tasks such as math problem solving.

The first component of failure mentioned above (i.e., how failure mechanisms are triggered) appears to occur quite differently in stereotype threat and performance under pressure. Specifically, performance pressure occurs when there are externally imposed consequences associated with poor performance. That is, there is an explicit expectation for high-level performance and, as a result, less-than-optimal performance ensues (Baumeister, 1984; Beilock & Carr, 2001). Stereotype threat, on the other hand, occurs because one has an awareness of one's social group membership and how members of those groups are expected to perform. Moreover, under stereotype threat, there is an explicit expectation for poor performance, which is the opposite of what occurs under pressure. Further, stereotype threat seems to be quite different from math anxiety as well. Although the former is thought to target those who are most invested in performing well and have the tools to do so (Spencer et al., 1999), individuals high in math anxiety often do not believe that they have the ability to succeed (Ashcraft & Kirk, 2001). Thus, from the standpoint of understanding the types of environmental triggers of failure and those individuals most likely to fail, stereotype threat, choking, and math anxiety seem quite different.

With respect to the second component mentioned above (i.e., how failure mechanisms operate), some similarities between choking, stereotype threat, and math anxiety exist. Pressure, stereotype threat, and even test anxiety have been shown to reduce the working memory capacity needed for task performance (Ashcraft & Kirk, 2001; Beilock et al., 2004; Beilock & Carr, 2005)—similar to our conclusions about stereotype threat in the current work. However, despite this similarity, the previously mentioned situations are often postulated to exert their impact via heightened levels of state anxiety (and are assessed by using state anxiety measures, e.g., Beilock et al., 2004; Tohill & Holyoak, 2000). In contrast, stereotype threat research has found weak relations between anxiety and impaired performance under stereotype threat (for a review, see Cadinu et al. 2005; also Schmader & Johns, 2003). The current work is no exception.

Moreover, the finding that stereotype threat can spill over and implicate tasks unrelated to the activated performance stereotype seems unique to the stereotype threat phenomenon. Performance pressure and test anxiety are triggered with respect to discrete events and their consequences (e.g., failing a test). In contrast, stereotype threat occurs when a pervasive stereotype is activated—a stereotype one might still view as threatening long after the specific performance situation one is in has ceased. Thus, the fact that a stereotype in one domain can have consequences for performance in another area provides insight into skill failure in a way that studying performance pressure or test anxiety cannot. An individual high in math anxiety may perform poorer than their nonmath-anxious counterpart on a math-based task, but this lower performance level does not appear on verbal tests (Ashcraft & Kirk, 2001). In contrast, in the current work we demonstrate that stereotype threat on a math task impacts performance on subsequent tasks unrelated to the stereotyped domain. As one might imagine, these findings have important implications for how overall performance may be affected by the ordering of sections on tests such as the SAT or the GRE.

Conclusions

In summary, the current work explored the cognitive mechanisms governing stereotype threat. In working memory intensive tasks such as mathematical problem solving, stereotype threat harms the cognitive system by co-opting working memory resources—and especially verbal resources—needed to perform certain types of math problems. This knowledge was used to devise a training regimen to alleviate these unwanted performance decrements as well as to predict when such performance failures would persist during the performance of subsequent tasks unrelated to the stereotyped task.

Our work not only demonstrates the value of examining underlying process as a means to gain a fuller theoretical understanding of stereotype threat but also provides an important theoretical bridge between work on stereotype threat (e.g., Steele, 1997; Steele et al., 2002; Wheeler & Petty, 2001) and research in cognitive psychology exploring test anxiety (e.g., Ashcraft & Kirk, 2001; Eysenck & Calvo, 1992) and performance pressure (Beilock & Carr, 2001; Beilock et al., 2004). Such cross-talk is vital for the development of comprehensive theories of failure that simultaneously take into account social and cognitive factors related to both the performer and the task being performed.

References

Aronson, J., Lustina, M., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition, 44,* 75–106.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130,* 224–237.

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Oxford University Press.

Baddeley, A. D. (1997). *Human memory: Theory and practice.* East Sussex, England: Psychology Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). New York: Cambridge University Press.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46,* 610–620.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74,* 1252–1265.

Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General, 130,* 701–725.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16,* 101–105.

Beilock, S. L., Jellison, W. A., Rydell, R. J., McConnell, A. R., & Carr, T. H. (2006). On the causal mechanisms of stereotype threat: Can skills that don't rely heavily on working memory still be threatened? *Personality & Social Psychology Bulletin, 32,* 1059–1071.

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General, 133,* 584–600.

Bogomolny, A. (1996). *Modular arithmetic*. Retrieved March 1, 2000, from http://www.cut-the-knot.com/blue/Modulo.shtml

Bosson, J. K., Haymovitz, E. L., & Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus nonverbal anxiety. *Journal of Experimental Social Psychology, 40,* 247–255.

Brosschot, J. F., & Thayer, J. F. (2003). Heart rate response is longer after negative emotions than after positive emotions. *International Journal of Psychophysiology, 50,* 181–187.

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16,* 572–578.

Carlson, R. A. (1997). *Experienced cognition.* Mahwah, NJ: Erlbaum.

Christoff, K., Ream, J. M., & Gabrieli, J. D. E. (2004). Neural basis of spontaneous thought processes. *Cortex, 40,* 623–630.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 25,* 257–271.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). New York: Cambridge University Press.

Darke, S. (1988). Effects of anxiety on inferential reasoning task performance. *Journal of Personality and Social Psychology, 55,* 499–505.

Davidson, R. J., Jackson, D. C., & Kalin, N. H. (2000). Emotion, plasticity, context, and regulation: Perspectives from affection neuroscience. *Psychological Bulletin, 126,* 890–909.

DeStefano, D., & LeFevre, J. A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16,* 353–386.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and function of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). New York: Cambridge University Press.

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion, 6,* 409–434.

Friedman, N. P., & Miyake, A. (2000). Differential roles for visuospatial and verbal working memory in situation model construction. *Journal of Experimental Psychology: General, 129,* 61–83.

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality & Social Psychology Bulletin, 28,* 659–670.

Gray, J. R. (2001). Emotional modulation of cognitive control: Approach–withdrawal states double-dissociate spatial from verbal two-back task performance. *Journal of Experimental Psychology: General, 130,* 436–452.

Gray, J. R., Braver, T. S., & Raichle, M. E. (2002). Integration of emotion and cognition in the lateral prefrontal cortex. *Proceedings of the National Academy of Sciences USA, 99,* 4115–4120.

Ikeda, M., Iwanaga, M., & Seiwa, H. (1996). Test anxiety and working memory systems. *Perceptual and Motor Skills, 82,* 1223–1231.

Johns, F., Schmader, T., & Martens, A. (2005). Knowing is half the battle—Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16,* 175–179.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217.

Kegel-Flom, P., & Didion, C. J. (1995, October 20). Women, math, and test scores. *Science, 270,* 364–365.

Klapp, S. T., Boches, C. A., Trabert, M. L., & Logan, G. D. (1991). Automatizing alphabet arithmetic: II. Are there practice effects after automaticity is achieved? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 196–209.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling working memory in a unified architecture: An ACT-R perspective. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 105–115). New York: Cambridge University Press.

Markham, R., & Darke, S. (1991). The effects of anxiety on verbal and spatial task performance. *Australian Journal of Psychology, 43,* 107–111.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42,* 236–243.

Martin, L. L., & Tesser, A. (1989). Toward a motivational and structural theory of ruminative thought. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 306–326). New York: Guilford Press.

Martin, L. L., & Tesser, A. (1996). Some ruminative thoughts. In R. S. Wyer, Jr., (Ed.), *Advances in social cognition* (Vol. 9, pp. 189–208). Mahwah, NJ: Erlbaum.

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General, 130,* 163–168.

Miyake, A., & Shah, P. (1999a). Toward unified theories of working memory: Emerging general consensus, unresolved theoretical issues, and future research directions. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 442–481). New York: Cambridge University Press.

Miyake, A., & Shah, P. (1999b). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126,* 247–259.

Rapee, R. M. (1993). The utilization of working memory by worry. *Behavioral Research Therapy, 31,* 617–620.

Richeson, J. A., & Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology, 88,* 934–947.

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

Schmeichel, B. J., Vohs, K. D., & Baumeister, R. F. (2003). Intellectual performance and ego depletion: Role of the self in logical reasoning and

other information processing. *Journal of Personality and Social Psychology, 85,* 33–46.

Seibt, B., & Forster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory focus. *Journal of Personality and Social Psychology, 87,* 38–56.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125,* 4–27.

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283,* 1657–1661.

Smith, J. L., & Johnson, C. S. (2006). A stereotype boost or choking under pressure? Positive gender stereotypes and men who are low in domain identification. *Basic and Applied Social Psychology, 28,* 51–63.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4–28.

Spielberger, C. C., Gorsuch, R. L., & Lushene, R. (1970). *State–Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 379–440). Amsterdam: Academic Press.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology, 77,* 1213–1227.

Teasdale, J. D., Dritschel, B. H., Taylor, M. J., Proctor, L., Lloyd, C. A., Nimmo-Smith, I., & Baddeley, A. D. (1995). Stimulus-independent thought depends on central executive resources. *Memory & Cognition, 23,* 551–559.

Tohill, J. M., & Holyoak, K. (2000). The impact of anxiety on analogical reasoning. *Thinking and Reasoning, 6,* 27–40.

Trbovich, P. L., & LeFevre, J. (2003). Phonological and visual working memory in mental addition. *Memory & Cognition, 31,* 738–745.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797–826.

Appendix

Stereotype Threat Manipulation

All participants read:

"We are interested in modular mathematics for a reason. As you probably know, math skills are crucial to performance in many important subjects in college. Yet surprisingly little is known about the mental processes underlying math ability. This research is aimed at better understanding what makes some people better at math than others."

Control group also read (Experiment 1 and Experiment 3B):

"Your performance on the math problems you are doing today will be compared to other students from across the nation."

Stereotype threat group (Experiment 1) and all participants prior to the stereotype threat block (Experiments 3–5) also read:

"As you also may know, at most schools male students outnumber female students in math majors and majors with math as a prerequisite, and there seems to be a growing gap in academic performance between these groups. A good deal of research indicates that males consistently score higher than females on standardized tests of math ability. But thus far, there is not a good explanation for this. The research you are participating in is aimed at better understanding these differences. Your performance on the math problems you are doing today will be compared to other students from across the nation. One specific question is whether males are superior at all types of math problems or only certain types."

Received February 28, 2005
Revision received September 26, 2006
Accepted September 26, 2006 ■