Research Report

When High-Powered People Fail

Working Memory and "Choking Under Pressure" in Math

Sian L. Beilock[1] and Thomas H. Carr[2]

[1]Miami University and [2]Michigan State University

ABSTRACT—*We examined the relation between pressure-induced performance decrements, or "choking under pressure," in mathematical problem solving and individual differences in working memory capacity. In cognitively based academic skills such as math, pressure is thought to harm performance by reducing the working memory capacity available for skill execution. Results demonstrated that only individuals high in working memory capacity were harmed by performance pressure, and, furthermore, these skill decrements were limited to math problems with the highest demands on working memory capacity. These findings suggest that performance pressure harms individuals most qualified to succeed by consuming the working memory capacity that they rely on for their superior performance.*

For many people, the desire to perform their best in academics is high. Consequences for suboptimal performance, especially in examinations, include poor evaluations by mentors, teachers, and peers; lost scholarships; and relinquished educational and employment opportunities. However, in comparison to research examining the cognitive processes underlying skill learning and execution (e.g., Anderson, 1993; Ericsson & Charness, 1994; Rosenbaum, Carlson, & Gilmore, 2001), little work has addressed the causal mechanisms by which high-stakes situations result in disappointing performances. Even less is known about the characteristics of those individuals most likely to experience unwanted skill failures.

Recently, researchers in cognitive and social psychology have begun to address these issues. Three recent studies have focused specifically on failure in mathematics. Ashcraft and Kirk (2001) examined how math anxiety undermines the performance of individuals who, in non-anxiety-provoking task domains, are highly competent. Schmader and Johns (2003) examined the cognitive mechanisms responsible for *stereotype threat* in math. Stereotype threat occurs when awareness of a negative stereotype about a social group in a particular task results in less-than-optimal performance by members of that group (Steele, 1997). And we (Beilock, Kulp, Holt, & Carr, 2004) explored the cognitive processes governing "choking under pressure." Choking, or performing more poorly than expected given one's skill, occurs in situations in which the desire for high-level performance is maximal (Beilock & Carr, 2001).

Surprisingly, these studies of diverse phenomena yielded similar conclusions concerning how suboptimal performance arises in mathematical problem solving. All involved working memory, a short-term memory system that maintains, in an active state, a limited amount of information with immediate relevance to the task at hand while preventing distractions from the environment and irrelevant thoughts (Hasher, Zacks, & Lustig, in press; Kane & Engle, 2000, 2002). If the ability of working memory to maintain task focus is disrupted, performance may suffer.

Ashcraft and Kirk (2001), and other anxiety researchers (Eysenck & Keane, 1990), have suggested that anxiety generates intrusive worries about the situation that occupy part of the working memory capacity normally devoted to skill execution. Moreover, research by Gray (2001; Gray, Braver, & Raichle, 2002) indicates a "double whammy," because anxiety is an unpleasant emotion, and unpleasant emotional states reduce the working memory capacity available for any verbal information, whether necessary task information or situational worries. Schmader and Johns (2003) argued that stereotype threat also interferes with performance by consuming or reducing the working memory capacity that individuals need to perform successfully. Finally, we (Beilock et al., 2004) found support for distraction theories of choking, according to which, like anxiety, pressure creates mental distractions that compete for and reduce working memory capacity that would otherwise be allocated to skill execution.

Together, this work suggests that compromises of working memory cause failure in tasks that rely heavily on this system. However, knowledge of the causal mechanisms governing

Address correspondence to Sian L. Beilock, Department of Psychology, Miami University, Room 202, Benton Hall, Oxford, OH 45056; e-mail: beilocsl@muohio.edu.

suboptimal performance is only part of the key to understanding failure. To truly understand unwanted skill decrements, and to engineer training regimens to alleviate them, one must also identify characteristics of those individuals most likely to fail.

Toward this end, the current experiment explored how individual differences in working memory capacity might be involved in susceptibility to choking under pressure in mathematical problem solving. An obvious hypothesis is that individuals low in working memory capacity (LWMs) are more prone to choke under pressure than are individuals high in working memory capacity (HWMs) because LWMs have limited capacity to compute problem solutions to begin with. Consequently, pressure-induced consumption of working memory might shrink available capacity below the minimum needed to solve a problem successfully.

However, another possibility exists: HWMs might be more prone to pressure-induced failure than are LWMs. Suppose HWMs rely more than do LWMs on strategies that load working memory during problem solution—"if you've got it, flaunt it." If so, under normal conditions, HWMs should perform better than LWMs on difficult tasks, as HWMs should have more resources to devote to problem solving. However, HWMs' usual working memory advantage may be just what makes them susceptible to failure when pressure is added, if pressure-induced consumption of working memory denies them the capacity they normally rely on to produce their superior performance. A similar argument has been made concerning working memory and the performance of attention-demanding verbal fluency and proactive-interference tasks. Under single-task conditions, HWMs outperform LWMs on such tasks. However, adding a secondary task essentially makes HWMs perform like LWMs by reducing the capacity that HWMs normally rely on to deal with the extra attention demands of difficult tasks (Kane & Engle, 2000, 2002).

If pressure and anxiety target individuals high in working memory capacity, this would carry significant implications for interpreting performance in high-pressure situations (e.g., college entrance exams). First, it would suggest that individuals most equipped to handle difficult situations that are working memory intensive (i.e., HWMs) are the ones most likely to "blow it" under pressure. Second, as working memory capacity is known to mediate and predict higher-level functions from comprehension to learning (Engle, Kane, & Tuholski, 1999), such results would call into question the ability of performance in high-pressure situations to differentiate persons most qualified to succeed from those with less capacity-related potential.

THE CURRENT EXPERIMENT

We chose Gauss's (1801) modular arithmetic (MA) task (cited in Bogomolny, 1996) to explore these hypotheses. The object of MA is to judge the truth value of problem statements such as "$51 \equiv 19 \pmod 4$." The problem is solved by subtracting the middle number from the first number (i.e., $51 - 19$) and then dividing this difference by the last number (i.e., $32 \div 4$). If the dividend is a whole number (here, 8), the statement is true. MA is similar to real-world math, as it is based on subtraction and division procedures. However, because MA is novel, even to most people highly experienced in math, it is advantageous as a laboratory task.

In the current study, individuals performed MA problems under both low-pressure and high-pressure conditions. The problems were manipulated to be either low or high in working memory demands. If pressure consumes the working memory capacity available for MA, then problems that depend heavily on working memory should suffer most when the problem solver is under pressure. Furthermore, if individual differences in working memory capacity are related to performance, this relationship should be most evident for MA problems that make the heaviest demands on working memory.

For present purposes, higher versus lower working memory demand was determined by whether the first step in solving the MA problem did or did not have a large number (> 20) or require a borrow operation. For example, "$5 \equiv 3 \pmod 2$" involves small numbers in the first step ($5 - 3$) and no borrow operation, so it was considered to have a low working memory demand. In contrast, "$45 \equiv 27 \pmod 4$)," involves both large numbers ($45 - 27$) and borrowing in the first step, so it was considered to have a higher working memory demand. Large numbers and borrow operations involve longer sequences of steps and require maintenance of more intermediate products, thereby placing heavy demands on working memory (Ashcraft, 1992; Ashcraft & Kirk, 2001).

METHOD

Participants

Data from 93 Michigan State University undergraduates were analyzed. Participants were divided into an LWM group ($n = 47$) and an HWM group ($n = 46$) using a median split of the average of their scores on two working memory tests: Turner and Engle's (1989) Operation Span (OSPAN) and a modified version of Daneman and Carpenter's (1980) Reading Span (RSPAN). The OSPAN involves solving a series of arithmetic equations while attempting to remember a list of unrelated words. Individuals are presented with one equation-word string at a time [e.g., "$(5 \times 2) - 2 = 8$? DOG"] on a computer and asked to verify aloud whether the equation is correct. They then read the word aloud. At the end of the series, they write down the sequence of words. The RSPAN involves reading a series of sentence-letter strings (e.g., "On warm sunny afternoons, I like to walk in the park. ? F"). Individuals read each sentence aloud, are asked to verify whether it makes sense, and then read the letter aloud. At the end of the series, they write down the sequence of letters. In both the OSPAN and the RSPAN, each series consists of two to five strings, and the order of string length is determined ran-

domly. Individuals are tested on three series of each length (12 total). OSPAN and RSPAN scores (range: 0–42) consist of the total number of words or letters recalled on perfectly recalled trials.

Span scores averaged across the two tests ranged from 2 through 32 (LWMs: $M = 9.76$, $SE = 0.42$; HWMs: $M = 21.07$, $SE = 0.63$). Three more participants were tested but not included because their accuracy in the arithmetic or sentence-verification portions of the span tests was less than 80%, suggesting they did not perform the test successfully. However, a reanalysis including their data did not change the results in any way.

Procedure

Participants were tested individually on a computer. They were instructed to judge MA problems as quickly as possible without sacrificing accuracy, pressing the "T" or "F" key to indicate whether each problem was true or false, respectively.

Each trial began with a 500-ms fixation point at the center of the screen. It was immediately replaced by an MA problem that remained on screen until the participant responded. After the response, the word "Correct" or "Incorrect" appeared for 1,000 ms, providing feedback. The screen then went blank for a 1,000-ms intertrial interval.

Individuals first performed three low-demand [e.g., $7 \equiv 2 \pmod 5$] and three high-demand [e.g., $44 \equiv 28 \pmod 7$] practice problems, which were presented in a different random order to each participant.

Participants then completed a 24-problem low-pressure test and a 24-problem high-pressure test. The problems in each test were presented in a different random order to each participant. Each problem appeared only once for each participant, in either the low-pressure or the high-pressure test, and the problems in the two tests were counterbalanced across participants.[1] Within each test, there were 12 low-demand and 12 high-demand problems. Half the problems within each demand level were true. The low-pressure test was described as more practice. Following this test, participants were given a scenario designed to create a high-pressure environment by involving sources of pressure commonly seen in the real world (monetary incentives, peer pressure, and social evaluation).

Participants were informed that the computer used reaction time (RT) and accuracy, equally, to compute an MA score. They were told that if they could improve their MA score by 20% relative to the preceding practice trials, they would receive $5. Each participant was further informed that obtaining the award required "team effort": He or she had been randomly paired with another individual, and for either person to receive $5, both members of the pair had to improve. Next, the participant was told that this partner had already completed the experiment and had improved by 20%. If the participant improved by 20%, both the participant and the partner would receive $5. However, if the participant did not improve by the required amount, neither individual would receive money. Finally, the participant was told that his or her performance would be videotaped so that local math teachers and professors could examine his or her performance on this new task. The experimenter set up the video camera (0.61 m to the right of the participant) to record the participant and the computer screen. The participant then completed the block of 24 MA problems.

This scenario has been repeatedly demonstrated to induce performance decrements across different skills, as well as to increase feelings of pressure and anxiety. These increased perceptions of pressure and anxiety do not differ as a function of math ability or performance under low-pressure conditions (Beilock & Carr, 2001; Beilock et al., 2004), so these factors were not likely to be confounded with response to pressure.

Following the MA tests, participants completed a paper-and-pencil division task and a subtraction and multiplication task. They were informed that these tasks were independent of the MA task. They were told that they should perform the tasks as quickly and accurately as possible, but that they were not expected to finish. These tasks served as fillers without pressure, designed to diminish any residual feelings created by the high-pressure situation prior to the OSPAN and RSPAN, which were administered next. For these, participants were simply informed to try their best. Upon completion of the experiment, all participants were given $5 and debriefed.

RESULTS

MA accuracy was examined in a 2 (working memory group: LWM, HWM) × 2 (problem demand: low, high) × 2 (pressure: low, high) analysis of variance (ANOVA). A significant three-way interaction was obtained, $F(1, 91) = 6.32$, $p < .02$, $\eta_p^2 = .07$. As shown in the upper left graph in Figure 1, LWMs were not influenced by pressure. This was confirmed by a 2 (problem demand: low, high) × 2 (pressure: low, high) ANOVA, revealing a main effect of problem demand, $F(1, 46) = 137.13$, $p < .01$, $\eta_p^2 = .75$; no main effect of pressure, $F < 1$; and no interaction, $F(1, 46) = 1.70$, n.s.

In contrast, a similar ANOVA with HWMs (Fig. 1, upper right graph) revealed a significant Problem Demand × Pressure interaction, $F(1, 45) = 4.89$, $p < .04$, $\eta_p^2 = .10$. Although HWMs' performance on low-demand problems did not differ as a function of pressure, $t(45) = 0.18$, n.s., their performance on high-demand problems declined significantly on the high-pressure test, $t(45) = 2.36$, $p < .03$, $d = 0.39$. This result is consistent with the idea that pressure consumes the working memory that HWMs use for successful performance of the most difficult problems with high working memory demands. Indeed, the high-pressure situation completely eliminated the advantage that HWMs enjoyed over LWMs on the high-demand

[1]The order of equations in the low-pressure and high-pressure tests produced no significant main effects or interactions with working memory group.

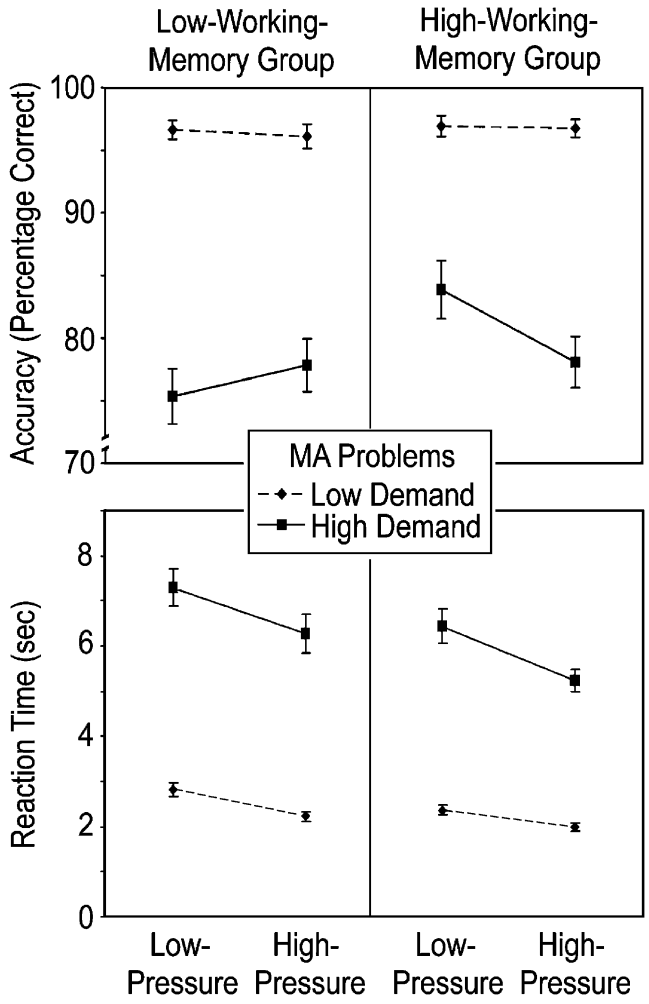

Fig. 1. Mean accuracy (upper graphs) and mean reaction time (lower graphs) for the group with low working memory capacity (left) and the group with high working memory capacity (right). Results are shown separately for low-demand and high-demand modular arithmetic (MA) problems in the low-pressure and high-pressure tests. Error bars represent standard errors.

problems in the low-pressure situation. This was confirmed by a separate 2 (working memory group: LWM, HWM) × 2 (pressure: low, high) ANOVA on accuracy on high-demand problems, which produced a significant interaction, $F(1, 91) = 5.97, p < .02, \eta_p^2 = .06$. In the absence of pressure, accuracy was significantly higher for the HWM than the LWM group, $t(91) = 2.65$, $p < .01$, $d = 0.55$. With pressure applied, this difference disappeared, $F < 1$.

RTs were analyzed for correct problems (Fig. 1, lower graphs). A three-factor ANOVA demonstrated that HWMs' RTs were faster than LWMs', $F(1, 91) = 4.22$, $p < .05$, $\eta_p^2 = .04$; RTs were slower for the high-demand problems than the low-demand problems, $F(1, 91) = 422.04$, $p < .01$, $\eta_p^2 = .82$; and all participants, regardless of working memory group, were slower in the low-pressure than the high-pressure test, $F(1, 91) = 92.50$, $p < .01$, $\eta_p^2 = .50$. There were no interactions involving working memory group and pressure. The absence of any interaction with working memory group in the RT results suggests that the differences in accuracy as a function of working memory were not produced by a speed-accuracy trade-off.

DISCUSSION

We examined the relation between pressure-induced performance decrements in mathematical problem solving and individual differences in working memory capacity. Decrements under pressure were limited to problems that made the largest demands on working memory—as one might expect. Surprisingly, however, only individuals high in working memory capacity showed these decrements. Individuals lower in working memory capacity performed less well on the high-demand problems in the absence of pressure, but when pressure was applied, LWM's disadvantage disappeared because their level of achievement did not decline under pressure. Working memory is at heart the ability to focus attention on a central task and execute its required operations while inhibiting irrelevant information (Hasher et al., in press; Kane & Engle, 2000, 2002). Under normal conditions, HWMs outperform LWMs because they have superior attentional allocation capacities of these types. When such attentional capacity is compromised, however, HWMs' advantage disappears.

The idea that pressure specifically targets individuals who have high working memory capacity carries implications for interpreting performance in real-world high-pressure situations. There is considerable debate concerning the ability of high-pressure tests (e.g., SAT, or Scholastic Assessment Test; GRE, or Graduate Record Examination) to predict future academic performance (Atkinson, 2001; Kuncel, Hezlett, & Ones, 2001; Sternberg & Williams, 1997). The current work adds to this debate by demonstrating, ironically, that the individuals most likely to fail under pressure are those who, in the absence of pressure, have the highest capacity for success.

REFERENCES

Anderson, J.R. (1993). *Rules of mind*. Hillsdale, NJ: Erlbaum.

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.

Atkinson, R.C. (2001, February). *Standardized tests and access to American universities*. The 2001 Robert H. Atwell Distinguished Lecture, presented at the annual meeting of the American Council on Education, Washington, DC.

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., 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

Daneman, M., & Carpenter, P.A. (1980). Individual-differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

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.

Ericsson, K.A., & Charness, N. (1994). Expert performance—its structure and acquisition. *American Psychologist*, *49*, 725–747.

Eysenck, M.W., & Keane, M.T. (1990). *Cognitive psychology: A student's handbook*. Hillsdale, NJ: Erlbaum.

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.

Hasher, L., Zacks, R.T., & Lustig, C. (in press). Variation in working memory due to aging and circadian arousal. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory*. New York: Oxford University Press.

Kane, M.J., & Engle, R.W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 336–358.

Kane, M.J., & Engle, R.W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637–671.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *27*, 162–181.

Rosenbaum, D.A., Carlson, R.A., & Gilmore, R.O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology*, *52*, 453–470.

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

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

Sternberg, R.J., & Williams, W.M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? A case study. *American Psychologist*, *52*, 630–641.

Turner, M.L., & Engle, R.W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

(Received 1/26/04; Revision accepted 3/9/04)