Available online at www.sciencedirect.com
ScienceDirect
Methods 42 (2007) 58–67
METHODS
www.elsevier.com/locate/ymeth

Methodologies for examining problem solving success and failure ☆

Marci S. DeCaro [a], Mareike Wieth [b], Sian L. Beilock [c,*]

[a] *Department of Psychology, Miami University, USA*
[b] *Department of Psychology, Albion College, USA*
[c] *Department of Psychology, 5848 South University Avenue, The University of Chicago, Chicago, IL 60637, USA*

Accepted 11 December 2006

Abstract

When designing research to examine the variables underlying creative thinking and problem solving success, one must not only consider (a) the demands of the task being performed, but (b) the characteristics of the individual performing the task and (c) the constraints of the skill execution environment. In the current paper we describe methodologies that allow one to effectively study creative thinking by capturing interactions among the individual, task, and problem solving situation. In doing so, we demonstrate that the relation between executive functioning and problem solving success is not always as straightforward as one might initially believe.

Keywords: Working memory; Performance; Pressure; Individual differences; Problem solving; Creativity; Short term memory; Stress; Math

1. Introduction

What determines successful performance on problem solving tasks ranging from insightful discovery to mathematical computation? In the current paper we explore methodologies used to investigate creativity in the problem solving domain—methodologies that take into account how individual differences in the performer, variations in the performance situation, and the demands of the task being performed carry implications for skill success and failure. By considering factors associated with the task, the performer, and the skill execution environment, we demonstrate that the relation between executive functioning and creative problem solving is not always as straightforward as one might initially believe.

Creativity has been broadly defined as the ability to produce original and appropriate problem solutions [1]. One specific way researchers have conceptualized creative thinking in the problem solving domain is in terms of cognitive flexibility. By flexibility we mean the ability to test multiple hypotheses or integrate numerous ideas, while filtering out unsuitable solutions, in order to arrive at an appropriate problem outcome [2]. Flexibility has also been conceptualized in terms of the ability to approach a problem in multiple ways, to develop new problem representations [3,4], and to come up with problem solving strategies that circumvent the impact of one's previous experience or tendency to solve a problem in a particular way [5].

Apparent in the above, multifacited, description of flexibility is the notion that creativity can be manifest in different types of problems in different ways. In the current paper we focus on two classes of problem solving tasks (well-structured and ill-structured problems [1]) thought to differ in the type of cognitive flexibility on which creative solutions depend, as a means to demonstrate that successful problem solving performance is a confluence of the individual, task, and performance environment.

Well-structured problems have clear initial or beginning states and clear goal states and are thus said to consist of a structured problem space [1]. Although not all well-structured problems are considered to be examples of creative problem solving, there are numerous examples of well-structured tasks that depend on the selection of one

☆ This research was supported by IES Grant R305H050004 and NSF Grant BCS-0601148 to Sian Beilock.
* Corresponding author.
E-mail address: beilock@uchicago.edu (S.L. Beilock).

1046-2023/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2006.12.006

solution path out of multiple possible options. These problems are said to rely on the ability to flexibly consider multiple ideas or algorithms simultaneously. Common move problems such as the "Hobbits and Orcs" have been shown to require such flexible thinking [1]. The goal of these problems is to take a number of reasoned steps in order to move six creatures (e.g., 3 hobbits and 3 orcs) across a river, while at the same time abiding by specific constraints concerning which creatures can be on the same side of the river at a given time. Specifically, all six creatures must cross to the opposite river bank using a rowboat with a maximum capacity of two. If the orcs on either side of the river outnumber the hobbits, the orcs will eat the hobbits. This task can be solved in a minimum of 11 steps and requires flexible thinking not only in planning and organizing these steps but also because individuals often find themselves in situations in which they have to make a move in the short-run that gets them farther away from their goal state.

Ill-structured problems generally do not have a clear initial problem state and thus the optimal solution space must be discovered [1]. Insight problems are commonly characterized as ill-structured in that these problems involve seeing information in novel ways in order to find the optimal path to solution. For example, the common two-string problem [6], in which one is asked to decide how to tie two strings hanging at opposite ends of the room together, involves using an object (e.g., a screwdriver) for a function not usually associated with it (e.g., the screwdriver is tied to one string and swung as a pendulum in order to afford the grasping of both strings to tie them together). Solving this task requires representing the use of familiar objects in uncommon ways. These sorts of insights are often associated with creative discoveries such as Archimedes' sudden realization to use water displacement as a tool to measure the king's crown after settling down in his bath [7] or Newton's theorization of gravity after watching an apple fall from a tree [2].

As one might imagine from the problem descriptions above, well-defined and ill-defined problems often rely on different types of cognitive flexibility for successful execution—flexibility in systematically considering and evaluating various problem solving strategies within a structured problem space versus flexibly adopting different initial approaches to a problem from the outset. As such, this problem distinction provides a useful framework within which to study the cognitive processes underlying creativity. It should be noted, however, that rather than conceiving of these problems as completely dichotomous, it may best to think of these problems as lying along a continuum [1]. After all, individuals can approach well-defined problems with a maladaptive problem representation, and ill-defined problems may sometimes involve intense concentration and systematic hypothesis testing. Nonetheless, for the purposes of the present work, the well-defined vs. ill-defined problem distinction provides a useful test bed for exploring creative thinking, its relation to individual differences in cognitive ability, and the impact of demanding environmental conditions on problem solving success.

We begin with a brief review and detailed methodological account of one individual difference variable hypothesized to be related to creative thinking—working memory capacity [2]. Working memory is believed to be an important cognitive construct underlying both skill learning and performance as it relates to several higher-level cognitive functions (e.g., general intellectual ability, reasoning, and analytic skill [8–10]). It is a common belief that general intelligence is highly associated with creative ability [2]. Thus, one might imagine that the higher one's working memory, the better his or her performance on the types of problem solving tasks outlined above. However, as will be seen below, these relations are not always as straightforward as one might initially imagine.

We next outline methods used to investigate how individual differences in working memory relate to problem solving performance as a function of the skill execution environment. Specifically, we describe recent research that has used the impact of high-stakes testing situations on performance as a tool to shed light on the link between working memory and problem solving success. As a preview, this work demonstrates that the widely assumed positive relationship between working memory, creative thinking, and performance can be reversed depending on the type of task being performed and the skill execution environment. As such, we conclude by discussing the import of taking into account interactions of the individual, task, and performance environment when developing methodologies to investigate creativity.

2. Working memory

Working memory (WM) can be described 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 [11]. Working memory has been distinguished from short-term memory capacity which is primarily thought to reflect information storage [12]. Primary functions of WM include retrieving, storing, and manipulating skill-relevant information as well as inhibiting information irrelevant to the task at hand [13,14]. Because short-term memory stores are best maintained when potential distractors are inhibited from entering, many view working memory primarily as an executive attention construct [15–17]. Consistent with this idea, the prefrontal cortex, a key structure in the distributed WM system, appears to be important in maintaining task information in an active state by reducing interference from competing sources both within and outside one's memory system [15,17,18].

Although working memory is conceptualized as a cognitive construct supporting complex thinking, it can also be thought of as an individual differences variable—meaning some people have more of this capacity and some have less. With this variability in mind, measures have been developed to assess individual differences in WM capacity as a means to capture the role of this important cognitive construct in performance. Broadly speaking, when representing WM as an individual difference variable, there are three major conclusions one can come to regarding the relationship between working memory and problem solving performance: (1) WM span positively correlates with task performance, implicating a potentially crucial WM mechanism underlying successful performance; (2) WM span and task performance are unrelated, indicating that successful performance does not rely on WM functions; or (3) WM span and task performance are negatively correlated, suggesting that successful performance may operate best outside of WM. This third option implies that the allocation of explicit attentional resources to performance may actually result in a less-than-optimal outcome. Thus, examining the relationship between available WM capacity and performance can offer insight into when WM resources may be positively correlated with the creative thinking necessary for successful outcomes on different problem solving tasks and when the opposite may occur.

2.1. Assessing working memory

There are a number of measures that are used to capture the cognitive construct of working memory and a number of texts devoted exclusively to describing them (e.g., see [11,12]). Below we outline a few common WM measures used to date and provide a detailed methodology for their implementation. Although many of these tasks differ in terms of their specific processing demands (e.g., solving math problems, judging sentences for comprehension), they are thought to reflect a common WM construct [12] that correlates highly with measures of general intelligence and reasoning ability (e.g., fluid intelligence [16]).

Two common WM span tasks are the reading span and operation span tasks. These tasks have proven to be reliable and valid measures of WM capacity [12] and have been implemented in studies examining performance on tasks ranging from math problem solving to categorization (e.g., [19,20]). Of note, our use of these tasks is adapted from Engle and colleagues [12,21]. These tasks are currently available to download from their website (http://psychology.gatech.edu/renglelab/; for a user's guide, see [12]). Moreover, use of the measures described below is recommended for healthy young adults. As a result, modifications may be necessary with other subject populations (e.g., older adults, children; see [12]).

2.1.1. Reading span

The reading span task (RSPAN; adapted from [22]) requires participants to process sentences while actively maintaining a small number of letters in memory. Participants are provided with an answer sheet comprised of 12 numbered lines with 5 blanks each and are instructed to memorize letters they see on the computer screen in addition to performing a second task between the presentation of each letter. The second task calls for judging the meaningfulness of unrelated sentences. For example, participants see sentence–letter displays such as "Whenever I drink the newspaper, I always get depressed. M." All sentences are syntactically correct, but many sentences, such as the preceding example, are not semantically correct. For each trial, participants view a sentence on the computer screen, read it out loud, and respond "yes" or "no" as to whether the sentence makes sense. The experimenter, seated beside the participant, marks down the participant response on an answer sheet. The participant then reads out loud the capitalized letter positioned to the right of the sentence. Immediately after the letter is read, the experimenter advances the screen by pressing the spacebar on the computer keyboard, and another sentence appears on the screen.

Participants must hold the letter(s) from all preceding sentence-letter trials in memory while processing further sentences until they see a screen with a series of question marks (i.e., "???"). At this prompt, participants write down all the letters they remember from the preceding trials (ranging in number from 2 to 5) in order. Participants are told there is no penalty for guessing and they are given no time limit for this recall period. After individuals finish writing down the letters, the experimenter advances the screen to the next set of sentence–letter trials.

Participants are first given three practice trials of two sentence–letter trials each. Then participants complete 12 sets of 2–5 trials each, in a predetermined random order, with three sets of each trial length (42 trials in total). A participant's RSPAN score can either consist of the total number of correctly recalled letters or only the number of letters recalled from all correctly-recalled sets (i.e., if participants recall any letters within a set incorrectly, a score of 0 is assigned for that set). In addition, individuals who fail to perform the sentence task at a high level of accuracy are usually removed from the dataset under the guise that they are not allocating appropriate resources to the sentence comprehension task, which may have consequences for letter recall.

2.1.2. Operation span

The operation span task (OSPAN [23]) is set up exactly the same as the RSPAN just described, with two exceptions. First, instead of processing sentences, participants read math operations of the type "IS $(6 \times 2) - 2 = 9$?" out loud and respond "yes" or "no" concerning whether the equation is correct. Second, instead of remembering letters, the OSPAN requires participants to remember simple words after responding to the equations.

2.1.3. Automated span tasks

Automated versions of the RSPAN and OSPAN were recently developed as alternatives to the traditional WM span measures described above. Because these measures are automated, experimenters can administer multiple sessions simultaneously and scores are conveniently generated after the session is complete. The automated OSPAN has been shown to be both reliable and valid and, although it only correlates moderately with the original OSPAN task [23], it shows the same general pattern of correlations with other measures of higher-order cognition as the paper and pencil OSPAN (e.g., Gf [21]). To the extent that one is interested in exploring relations between working memory and complex cognitive activities, this latter correlation is of utmost importance.

2.1.3.1. Automated reading span. In the automated reading span (Arspan [12]), participants perform three initial task blocks. In the first block, a short series of letters is presented on the computer screen, one after the other, for 800 ms each. After each set, participants are shown a 4×3 matrix of letters (F, H, J, K, L, N, P, Q, R, S, T, and Y) and, using the mouse to click a box beside each letter, are asked to recall the letters in the exact order in which they saw them. This recall stage is untimed, and participants are allowed to clear the screen and redo their responses as many times as necessary. After recall, participants receive feedback concerning the number of letters they recalled correctly in the current set.

The second block introduces the sentence processing task. Participants view a sentence (e.g., "The ranger told the hikers to watch out for snakes"), pressing the mouse button after reading it. Then participants see a screen with the question "Did the sentence make sense?" above "true" and "false" options, with the correct answer to be selected using the mouse cursor. Participants are asked to read the sentence as quickly as possible and to be as accurate as possible when responding to the question. Notably, the pace at which participants read the sentences during this practice block determines each individual's time limit during the test blocks (mean response time plus 2.5 SD).

In the third block, the letter span and sentence processing tasks are combined. Participants first view and respond to a sentence, then see the briefly presented letter. After a series of two of these sentence–letter trials, the recall screen appears and participants select the letters they remember in the correct order (3 sets total). Following recall, feedback concerning both the number of letters correctly selected and sentence accuracy is provided.

After these practice blocks, participants perform the experimental block of trials. Participants are instructed to keep their sentence accuracy above 85%, and a running accuracy total remains on the top right corner of the screen in red font. If individuals do not respond to the sentence before their allotted time runs out, that sentence is scored as incorrect. Three sets of each size (3–7 trials) are administered in random order, for a total of 75 sentence–letter trials.

After completion of the task, the experimenter can access the WM scores. Three main scores are of key interest. The Absolute Score is the sum of all perfectly recalled sets (i.e., the traditional scoring method). The Total Score is the sum of all correctly recalled letters. And the Reading Errors score sums the total errors on the sentence processing task (i.e., true/false errors and speed errors). It is standard practice to exclude data for participants who fail to achieve 85% accuracy on the sentence processing task in order to ensure that individuals are not devoting all their processing to remembering the letters [12,21].

2.1.3.2. Automated operation span. The automated operation span (Aospan [21]) follows the exact same structure as the Arspan, except that instead of processing sentences participants process simple math equations. Specifically, participants view an equation such as "$(1 \times 2) + 1 = ?$" and are instructed to press the mouse button after they have solved it. The equation is then replaced by a number, which participants judge as either the correct or incorrect equation answer by clicking "true" or "false" on the screen response box.

2.1.4. Two-back tasks

Although the RSPAN and OSPAN are commonly used measures of WM, there are several other methods used to assess this cognitive control construct as well. Many of these measures place a greater emphasis on domain-specific processes rather than the general executive attention processes believed to be assessed by the OSPAN and RSPAN tasks (see [12]). For example, research examining WM in problem solving has often utilized Baddeley's [24,25] multicomponent model of working memory as a guide. Baddeley's original model, in which different subsystems are thought to be devoted to different types of information, has 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 [26].

Because the verbal/visuospatial distinction has received a large amount of support in the human working memory literature [27,28], conceptualizing differences in problem solving in terms of verbal and visuospatial processing requirements provides a useful approach for examining skill execution. Moreover, functional neuroimaging evidence suggests that some components of working memory are hemispheric dependent—with verbal processes relying more on the left inferior frontal gyrus and spatial activities being more right lateralized [28–30]. Nonetheless, one should note that there is debate concerning whether working memory should be viewed primarily as a domain-general unitary system involved in executive-attention functions [31–33] or as a domain-specific system consisting of specialized components that handle specific types of information [24,25,34]. Individuals who argue for a domain-specific view do not deny that domain-general components exist [35]. 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 [32]. Thus, depending on the research question of interest, measures aimed at tapping domain-specific processing may be useful.

Below we describe one common method used to assess cognitive control as a function of the processing component involved—verbal and spatial two-back tasks [28,36]. In all cases, the presentation of the stimuli is the same, but the nature of the task (spatial vs. verbal) is manipulated by instructions presented to participants. Because the stimuli are held constant, with the processing demands of the task varying as a function of the instructions given, any performance differences are likely due to differences in domain-specific processing requirements rather than general difficulty differences across the two-back tasks [37].

In both the verbal and spatial two-back tasks, individuals are told to indicate whether a stimulus item presented on the current trial matches the item presented two trials previously by using either the "S" key (same stimuli) or the "D" key (different stimuli). The stimuli are 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 and against a background of random letters. On each trial, the target is presented for 500 ms, followed by a 2500 ms period during which only the background is shown. Thus, participants have 3000 ms to indicate their response before the next trial begins. If participants fail to respond, a tone is emitted and the trial is scored as an errant response.

Participants performing the verbal two-back task are asked to determine whether the letters of the current stimulus trial matched the letters presented two trials earlier (ignoring the physical location of those presentations). In contrast, individuals performing the spatial two-back task are asked to indicate whether the location of the current stimulus trial matched the same location as the stimulus presented two trials earlier (ignoring the letters presented). The first response trial occurs following the third stimulus presentation (where the stimulus is compared to the first stimulus presented). Participants are typically given an initial practice session to ensure they understood the task (repeated if necessary), followed by a critical session consisting of 100 response trials. In each session (practice and critical), 30% of the trials are "same" trials, and the remaining 70% are "different" trials.

The two-back tasks have been described as requiring seven operations: Encoding, storage, rehearsal, matching of current information to that stored in memory, temporal ordering (to maintain which item was stored two previously), inhibition, and response execution [38]. It should be noted that although we have conceptualized both the span tasks (i.e., the OSPAN and RSPAN) and the two-back tasks as capturing the general construct of WM, and despite the fact that both types of measures are commonly referred to as assessments of WM, further work is needed to clarify the extent to which these measures share variance in common with each other and with measures of general intellectual ability more broadly [12].

3. Performance under pressure

As mentioned above, working memory is thought to be involved in both the learning and performance of an array of complex problem solving tasks, ranging from reading comprehension [23], to logical reasoning [15], to mathematical problem solving [39]. Thus, in terms of studying the relation between working memory and creative thinking and problem solving, one approach would be to simply correlate individual differences in this cognitive construct with problem solving performance. An additional methodology, however, involves the instantiation of performance pressure during problem solving execution. High-pressure testing environments are thought to impact certain types of problem solving tasks by consuming the working memory resources individuals rely on to perform at an optimal level—and, as will be seen below, this outcome can differ across individuals [19]. Examining the relation between pressure, working memory, and performance outcomes then not only lends insight into the mechanisms underlying creative problem solving but also allows one to capture non-linear relations between working memory and creative thinking in the types of high-stakes situations under which tests of creativity often take place [40]. We focus on this method below.

The *distraction hypothesis* suggests that high-stakes testing leads to worries and thoughts about the situation and its consequences that reduce the cognitive capacity available for problem solving performance [39,19]. These worries are thought to rely most heavily on the phonological and executive control aspects of working memory believed to support inner speech and thinking in the service of complex cognitive activities [11,41]. Evidence in support of the distraction hypothesis comes from recent work employing the verbal and spatial two-back tasks described above. Beilock, Rydell, and McConnell [37] had individuals perform math problems under a high-stress condition followed by either a verbal or spatial two-back WM task. These tasks were matched for difficulty and appearance, differing most substantially in their reliance on either verbal or spatial WM processes [28]. Borrowing logic from the finding that depletion of resources in one task domain can carry over and impact performance on another task (e.g., [42]), if high-stress situations cause worries that consume verbal WM resources, and if this WM consumption does not immediately subside when performance on the math task is finished, then individuals should perform more poorly on a verbal (relative to a spatial) two-back task following a high-stress testing situation. In essence, test-induced worries may "spill over" onto other tasks that use the same processing resources but are not implicated by the high-stress situation. This is exactly what occurred.

Thus, high-stakes situations appear to impact the working memory resources available for problem solving performance via worries about the situation and its consequences. Prior to exploring the relation between WM and high-stakes execution any further however, we first turn to a detailed overview of one methodology used to create one type of high-stakes situation in the laboratory. We believe that this methodology provides much needed insight into problem solving success and failure and the relation between creative thinking and working memory in the types of high-stakes testing situations commonly encountered in the real world.

3.1. Implementing performance pressure

One high-pressure scenario we have consistently employed in our laboratory involves several sources of pressure that exist across skill domains—monetary incentives, peer pressure, and social evaluation. Although it is an empirical question as to exactly how these different sources of pressure exert their influence, the goal of our work is to create a high-stakes situation in the laboratory. Thus, we use a pressure scenario that incorporates as many components of high-pressure performance as possible. In athletics for example, performance is frequently scrutinized by others, there are often monetary consequences for winning and losing, and team success depends on the performance of individual athletes—which may generate peer pressure to perform at an optimal level. In more academic arenas, monetary consequences for test performance are manifested in terms of scholarships and future educational opportunities, and social evaluation of performance comes from mentors, teachers, and peers.

In terms of the specific procedure, participants first complete a set of problems (e.g., mathematical computations or insight problems) in a low-pressure condition, followed by another set of problems after a high-pressure scenario has been introduced. In our pressure manipulation, participants are given the following scenario (or a close adaptation thereof):

> Okay, you are probably wondering why you are doing these problems. As you have been going through the experiment, the computer has been keeping track of both your reaction time and accuracy and has been calculating a score for you based on these two components. In the next set of problems if you can *improve* your score by 20%, so if you can improve both your reaction time and accuracy by 20%, then at the end of the experiment, in addition to your experimental credit, we will give you [some money (eg., $10)].
>
> But, there is a catch. What we are interested in, in this experiment, is teamwork and how people work together. So, as part of this experiment you have been paired with another participant. In order for you to receive the [$10], not only do *you* have to improve by 20%, but the person that you have been paired with has to improve as well. So it's a "team effort." Now, I can tell you that you are actually the second of the pair. Your partner went through the experiment this morning and did improve by 20%. So, if you can improve now, you will get [$10] and so will your partner. But if not, they do not get the money and neither do you.
>
> Finally, during this next set of problems your performance is going to be videotaped. We are interested in how people perform on this new type of math task. Some professors and students at [university], and math teachers in the area will be watching the tapes to see how people are performing. Ok, I am going to set up the video camera now and then we can begin. [Experimenter sets up camera on a tripod, about 1.5 m to the right of participants, such that both their face and the computer screen or paper is in view]. Ok, do you see the red light? So again, if you can improve your reaction time and accuracy for solving the problems by 20% in the next set of problems, you will get [$10] and so will your partner. But if you don't improve, they don't get the money and neither do you. You can press the [key] to continue. [Experimenter stands behind camera].

After the problem set is completed, the experimenter turns off the camera and faces it away from the participant and tells the participant he or she is doing so and that the experimenter will inform the participant of their performance at the end of the experiment.

As a manipulation check, and to ensure that some individuals are not differentially impacted by the pressure manipulation than others, participants immediately complete a set of questionnaires designed to assess their feelings of performance pressure and state anxiety following performance under pressure. The first questionnaire, the State Form of the State-Trait Anxiety Inventory (STAI [43]), is a common measure of state anxiety consisting of 20 questions designed to assess individuals' feelings at a particular moment in time. Individuals are instructed to assign a value to questions such as, "I feel calm" and "I feel at ease" on a 4-point scale ranging from 1 (not at all) to 4 (very much so). Participants are then given two additional questions concerning their perceptions of performance under pressure [39]. Individuals rate (1) in comparison to the meaningful activities they have performed in the last week, how important they felt it was for them to perform at a high level on the problems (in the most recent block), ranging from 1 (not at all important to me) to 7 (extremely important to me), and (2) how much pressure they felt to perform at a high level, ranging from 1 (very little performance pressure) to 7 (extreme performance pressure).

We have found the above mentioned pressure scenario to be highly effective in producing feelings of performance pressure and anxiety in participants [19,39,44]. Furthermore, these feelings do not appear to differ as a function of problem solving ability or working memory capacity. Thus, susceptibility to feelings of pressure does not appear to be confounded with our key variables of interest.

Of note, when administering pressure within-subjects, the order of pressure conditions is typically not counterbalanced. Administering the low-pressure condition before the high-pressure condition is necessary in order to maintain a baseline, low-pressure measure of performance before pressure effects can take hold [19,39,44,45]. That is, it is hard to create a low-pressure situation once individuals have already completed a high-pressure condition. If condition order must be counterbalanced however, it is possible to have participants complete a filler task between high- and low-pressure blocks in an attempt to dissipate feelings of pressure and anxiety (e.g., [46]).

Although a lack of counterbalancing may create some concern, several factors typically alleviate the possibility that order effects are driving any obtained results. First, having more time with a task facilitates learning and performance [44,47]. Thus, if anything, performance on a second block of problems following performance pressure should be better than performance on a first problem block prior to a pressure manipulation. This practice effect works against the possibility of finding pressure-induced performance decrements. Second, the main contrast of interest is often consistent across the ordering of the pressure conditions. For example, if one is comparing the performance of individuals high in WM capacity to performance of individuals lower in WM under pressure, the comparison of interest is across WM groups. If order effects are at play, then both groups should be likewise affected. However, should the predicted relation be found between pressure condition and an independent variable of interest, then it would be difficult to explain simply by a main effect of pressure condition order.

3.2. Working memory, well-structured problems, and performance failure under pressure

As mentioned above, several studies have demonstrated that in cognitively-based problem solving tasks, high-pressure environments consume the WM resources that individuals might otherwise devote to task execution (see [39]). Thus, to the extent that well-structured problem solving tasks—in which individuals must flexibly consider a series of paths to reach a final goal state—rely on working memory capacity for successful execution, performance may suffer under stress.

Moreover, given the fact that individuals vary in the WM capacity they bring to a problem solving situation, the impact of pressure on problem performance may differ as a function of individual differences working memory. One obvious possibility is that individuals low in WM (low WMs) will be most prone to pressure-induced failure on these types of well-structured tasks. Such individuals have limited capacity to perform difficult tasks to begin with, and thus performance pressure may serve to diminish what few resources they have. However, another possibility is that individuals higher in WM (high WMs) are more prone to failure. Under normal, low-pressure conditions, high WMs should perform better than low WMs as they have more capacity to maintain and manipulate relevant aspects of the problems they are presented with. However, high WMs' usual working memory advantage may be just what makes them susceptible to failure when pressure is added if pressure denies high WMs the capacity they normally rely on to produce their superior performance. That is, performance pressure may make high WMs look like their low capacity counterparts—an unwanted outcome if the goal of a high-pressure test is to highlight performance differences among individuals with more or less capacity.

To test these competing predictions, Beilock and Carr [19] had individuals low and high in WM capacity perform a difficult math task under both low- and high-pressure conditions. Participants were divided into a low WM group and a high WM group based on a median split of the average of their scores on the OSPAN and RSPAN working memory tests (described in the working memory methods section above). Working memory was treated as a general executive attention construct in this work, as a means to assess how general capacity differences might interact with the demands of the task being performed and the performance environment to produce skill success and failure.

High WMs outperformed low WMs under low-pressure practice conditions. Such a result is not surprising given that high WMs should be better able to allocate the attentional resources needed to compute demanding well-structured problem solutions. However, when placed in a high-pressure testing situation, those highest in WM were the most likely to fail. In fact, high WMs' performance fell to the level of low WMs when the pressure was on. Low WMs' performance did not suffer under pressure. Similar results have been found using Raven's Standard Progressive Matrices as a test bed [48]—a task commonly used as a measure of creative thinking and reasoning. In this task, individuals are presented with increasingly difficult patterns that contain one missing segment and, given a list of possible options, are asked to choose the segment that best completes the pattern. Consistent with Beilock and Carr's [19] results, high WMs' performance dropped to the level of low WMs under pressure. Low WMs' performance remained consistently low across both the low- and high-pressure testing conditions.

How can one account for the above mentioned pattern of performance? Beilock and DeCaro [49] have recently shed light on this question by prompting individuals to describe the steps and processes they used to solve a selection of math problems under both low- and high-pressure conditions. Gauss's mathematical problem solving task, *modular arithmetic* [50], was used as a test bed. This task involves judging the truth-value of equations such as "$34 = 18 \pmod 4$." To do this task, the problem's middle number is subtracted from the first number (i.e., $34 - 18$) and this difference is then divided by the last number (i.e., $16 \div 4$). If the dividend is a whole number (here, 4), the statement is true. Although this task is based on common subtraction and division procedures, there are shortcut strategies that can be employed to derive the correct answer (some of the time) without requiring a multi-step problem solving algorithm. For example, if one concludes that all problems with even numbers are true, this shortcut strategy will produce the correct answer on some trials (as in the previous example), but not always (e.g., "$52 = 16 \pmod 8$"). Such shortcut strategies circumvent the need to maintain and manipulate intermediate problem steps on-line in working memory and thus should not be harmed by high-pressure situations that co-opt WM resources. However, because these shortcuts do not involve the computation of specific problem steps, they result in less accurate problem solving performance overall.

Under low-pressure conditions, high WMs reported relying on more working memory-intensive algorithms to solve the math problems (e.g., "I subtracted the numbers using a borrow operation and then divided"), whereas low WMs reported implementing more shortcut strategies (e.g., "If the numbers were all even, I assumed the answer to be true"). Under high-pressure testing conditions, both low and high WMs relied more on shortcut strategies to solve the math problems. Thus, rather than simply disrupting a complex problem solving procedure, performance pressure appears to have altered the problem solving approaches of high WM individuals—prompting them to rely on the less accurate, yet less WM-intensive, shortcut strategies normally employed by low WM individuals. Because low WMs were relying on simple shortcut strategies to begin with, their performance was not impacted by stress.

As just described, WM resources can aid problem solving by allowing a focused search for problem solutions while inhibiting irrelevant information. To the extent that creative problem solving tasks benefit from such resources, WM capacity may be important for expressing creativity. Indeed, it has been suggested that WM fosters cognitive flexibility, in that it supports strategic planning, access to long-term memory, and abstract thinking [2]. WM has even been posed as a prerequisite for creative thinking, because it underlies the ability to maintain and integrate information relevant for solving a problem [2,51]. Notably, however, this positive relation between working memory and performance in well-structured problem solving tasks is not always apparent. As seen above, working memory does not necessarily foster performance success in high-stress situations—especially to the extent that performance pressure most strongly impacts those individuals who rely most on WM for their superior performance. Moreover, as will be seen below, there are also conditions and problem solving tasks under which the positive relation between working memory and performance is not only absent, but can even be reversed.

3.3. Working memory, ill-structured problems, and performance success under pressure

A growing number of studies have demonstrated that there are tasks for which depleted WM resources actually *benefits* performance. For example, when learning stable and unchanging correlations between two events, low WMs have been shown to be more accurate than high WMs [52]. Modeling performance in ACT-R, Gaissmaier et al. [53], demonstrated that this pattern of data was due to high span individuals' tendency to employ complex hypothesis testing (e.g., probability matching), a strategy that produces poorer performance than the simpler strategy of always predicting that the most frequently occurring event will happen next (e.g., maximizing) [54].

Tasks for which an optimal solution path is often not readily apparent and instead relies on a novel or insightful way of representing a problem can be roughly categorized as *ill-structured problems*. To the extent that ill-structured tasks are best performed without explicit hypothesis testing, and performance pressure impinges on the controlled attention abilities that support hypothesis generation, then it seems possible that high-pressure environments may actually benefit execution—at least for those individuals prone to employ a complex solution that may misrepresent the problem in the first place (i.e., high WM individuals). To test this idea, Beilock and DeCaro [49] asked individuals to solve Luchins's water jug task [55]. In this task, the goal is to derive a mathematical formula to obtain a specified goal quantity of water with jugs of various capacities, using the *simplest* strategy possible. Participants were given six problems in total. The first three problems were solvable only by using a difficult, WM-demanding strategy (i.e., $B - A - 2C$). But the last three problems were solvable in either of two ways: (1) implementing the same difficult formula as used in the first problems, or (2) using a much simpler equation (i.e., $A - C$ or $A + C$). Of interest is whether individuals switch to the simpler shortcut strategy when available or whether they continue to use the demanding formula.

Under low-pressure conditions, high WMs were more likely to continue with the WM-intensive algorithm, whereas low WMs switched to the simpler solution when available. When individuals persist in using the difficult strategy in lieu of the simpler one, they are said to exhibit mental set. Mental set is typically viewed as a negative artifact of previous experience that leads individuals to focus exclusively on a narrow range of problem solving strategies and thus become oblivious to better solutions that may arise [5]. Because using less effort to produce a correct answer is more economical, and individuals were instructed to use the simplest strategy possible, the tendency to break mental set can be considered success on this task. Thus, low WMs outperformed high WMs, demonstrating that at times having greater ability to control attention can lead high WMs to miss more optimal creative problem solving approaches. Under high pressure however, high WMs switched to the simpler approach, equating their success to that of low WMs.

How can the above result be explained? Consistent with the view that attentional control abilities are a key component of WM capacity [15], high WMs may be especially good at focusing their attention on select task properties and ignoring others, whereas low WMs may not be able to allocate attentional resources solely to one task approach. As a result, low WMs may actually be better able to detect alternate problem solutions. Support for this notion comes from Conway et al.'s [56] examination of the dichotic listening performance of individuals lower and higher in working memory. Individuals were instructed to listen to a message in one ear and ignore a message in the other ear in which their name was sometimes mentioned. It was found that low WMs were more likely to notice their name in the unattended ear than high WMs, suggesting that the former group was allocating attention to information both focal and disparate to the task at hand. This inability to effectively hone one's attentional resources may help lower WM individuals find an easier way out of the water jug task and may also be adopted by higher WM individuals once pressure co-opts the resources they would normally use to allocate attention solely to one task approach.

As discussed previously, creative thinking is commonly portrayed as an ability to generate relevant problem solving techniques in a focused manner. In this sense, creativity seems more attainable for individuals with greater available WM resources. However, this greater capacity to focus attention by inhibiting irrelevant information necessarily limits the solution space as well [2]. Indeed, showing that WM does not always correlate with optimal performance (e.g., in tasks where explicit hypothesis testing is not helpful; see also [7]) perhaps warrants another view of creative problem solving. Specifically, for certain problem types (e.g., ill-structured problems), creative problem solving may be most adept when attention is defocused. A more diffuse attentional focus enables the solver to integrate presumably irrelevant information with task information. Indeed, the creative process has often been described as automatic and effortless, spurred even by states of altered consciousness such as commonly associated with daydreaming, substance use, or mental illness [2]. By implication then, perhaps creative performance on a task best approached by exploring even seemingly tangential ideas will prosper in a problem solving environment that limits WM involvement.

4. Conclusion

When designing studies to illuminate the variables responsible for creative problem solving performance, one must not only consider characteristics of the individual and the demands of the task being performed but the performance environment as well. In the current paper we aimed to portray a range of methodologies that can be used to investigate creativity in problem solving performance. This type of multifaceted approach to creative problem solving illuminates the notion that "success" on all problem solving tasks is not a simple linear factor of one's cognitive abilities. Indeed, the best and the brightest on some types of tasks (e.g., well-defined problem solving tasks involving explicit hypothesis testing) may be the most likely to fail at other tasks (e.g., ill-defined tasks best solved by abandoning complex and working memory-demanding algorithms). Moreover, to the extent that changes in the performance environment influence the cognitive resources at hand, those individuals who once performed at the top precisely because they relied on working memory resources for their superior execution may be most likely to fail when the pressure is on. This effect of situational variables may be of further import for creativity researchers when considering that common tests of creative thinking are situated within tight time constraints, personal desirability for good performance, and close examination by experimenters [40]—all aspects of the high pressure situation we create in our laboratory.

Further research taking constraints of the individual, task, and performance environment into consideration will not only advance our theoretical understanding of creativity in the problem solving domain but will also lead to the development of training regimens, assessment conditions, and educational settings that encourage the type of creative processing best suited for the task at hand. Such knowledge will benefit researchers, teachers, and performers alike and take a further step towards illuminating how the cognitive processes studied in the laboratory actually play out in the complexities of the real world.

References

[1] R.J. Sternberg, Cognitive Psychology, third ed., Wordworth, Thomson Learning, CA, 2003.
[2] A. Dietrich, Psychonomic Bulletin and Review 11 (2004) 1011–1026.
[3] F.G. Ashby, A.M. Isen, A.U. Turken, Psychological Review 106 (1999) 529–550.
[4] W.T. Maddox, G.C. Baldwin, A.B. Markman, Memory Cognition, in press.
[5] J. Wiley, Memory and Cognition 26 (1988) 716–730.
[6] W. Weiten, Psychology: Themes and Variations, third ed., Brooks/Cole, CA, 1995.
[7] J.W. Schooler, J. Melcher, in: S.M. Smith, T.B. Ward, R.A. Finke (Eds.), The Creative Cognition Approach, MIT Press, MA, 1995, pp. 97–133.
[8] P.C. Kyllonen, R.E. Christal, Intelligence 14 (1990) 389–433.
[9] R.W. Engle, S.W. Tuholski, J.E. Laughlin, A.R.A. Conway, Journal of Experimental Psychology: General 128 (1999) 309–331.
[10] D.Z. Hambrick, Memory and Cognition 31 (2003) 902–917.
[11] A. Miyake, P. Shah, Models of Working Memory: Mechanisms of Active Maintenance and Executive Control, University Press, New York, 1999.
[12] A.R.A. Conway, M.J. Kane, M.F. Bunting, D.Z. Hambrick, O. Wilhelm, R.W. Engle, Psychonomic Bulletin and Review 12 (2005) 769–786.
[13] V.M. Rosen, R.W. Engle, Journal of Experimental Psychology: General 126 (1997) 211–227.
[14] A. Shimamura, in: M.S. Gazzaniga (Ed.), The Cognitive Neurosciences, MIT Press, Cambridge, 1995, pp. 803–813.
[15] R.W. Engle, Current Directions in Psychological Science 11 (2002) 19–23.
[16] M.J. Kane, K.M. Bleckley, A.R.A. Conway, R.W. Engle, Journal of Experimental Psychology: General 130 (2001) 169–183.
[17] M.J. Kane, R.W. Engle, Psychonomic Bulletin and Review 9 (2002) 637–671.
[18] J.D. Cohen, T.S. Braver, R.C. O'Reilly, A Computational Approach to Prefrontal Cortex, Cognitive Control, and Schizophrenia: Recent Developments and Current Challenges, Oxford University Press, Oxford, 1998.
[19] S.L. Beilock, T.H. Carr, Psychological Science 16 (2005) 101–105.
[20] M.S. DeCaro, R.D. Thomas, S.L. Beilock, Working memory and categorization: sometimes less is more (under review).

[21] N. Unsworth, R.P. Heitz, J.C. Schrock, R.W. Engle, Behavior Research Methods 37 (2005) 498–505.
[22] M. Daneman, P.A. Carpenter, Journal of Verbal Learning and Verbal Behavior 19 (1980) 450–466.
[23] M.L. Turner, R.W. Engle, Journal of Memory and Language 28 (1989) 127–154.
[24] A.D. Baddeley, Working Memory, Oxford University Press, Oxford, 1986.
[25] A.D. Baddeley, R.H. Logie, in: [4].
[26] A.D. Baddeley, Trends in Cognitive Sciences 4 (2000) 417–423.
[27] D. DeStefano, J.A. LeFevre, European Journal of Cognitive Psychology 16 (2004) 353–386.
[28] J.R. Gray, Journal of Experimental Psychology: General 130 (2001) 436–452.
[29] M. D'Esposito, G.K. Aguirre, E. Zarahn, D. Ballard, R.K. Shin, J. Lease, Journal of Cognitive Brain Research 7 (1998) 1–13.
[30] E.E. Smith, J. Jonides, Science 283 (1999) 165.
[31] N. Cowan, in: [4].
[32] R.W. Engle, M.J. Kane, S.W. Tuholski, in: [4].
[33] M.C. Lovett, L.M. Reder, C. Lebiere, in: [4].
[34] N.P. Friedman, A. Miyake, Journal of Experimental Psychology: General 129 (2000) 61–83.
[35] A. Miyake, Journal of Experimental Psychology: General 130 (2001) 163–168.
[36] T.S. Braver, J.D. Cohen, L.E. Nystrom, J. Jonides, E.E. Smith, D.C. Noll, Neuroimage 5 (1997) 49–62.
[37] S.L. Beilock, R.J. Rydell, A.R. McConnell, Journal of Experimental Psychology: General, in press.
[38] J. Jonides, Journal of Cognitive Neuroscience 9 (1997) 462–475.
[39] S.L. Beilock, C.A. Kulp, L.E. Holt, T.H. Carr, Journal of Experimental Psychology: General 133 (2004) 584–600.
[40] E.M. Bowden, M. Jung-Beeman, J. Fleck, J. Kounios, Trends in Cognitive Science 9 (2005) 322–328.
[41] R. Carlson, Experienced Cognition, Lawrence Erlbaum Associates, Mahwah, 1997.
[42] R.F. Baumeister, E. Bratslavsky, M. Muraven, D.M. Tice, Journal of Personality and Social Psychology 80 (1998) 1252–1265.
[43] C.C. Spielberger, R.L. Gorsuch, R. Lushene, State-Trait Anxiety Inventory, Consulting Psychology Press, Palo Alto, 1970.
[44] S.L. Beilock, T.H. Carr, Journal of Experimental Psychology: General 130 (2001) 701–725.
[45] R. Gray, Journal of Experimental Psychology: Applied 10 (2004) 42–54.
[46] R.C. Jackson, K.J. Ashford, G. Norsworthy, Journal of Sport and Exercise Psychology 28 (2006) 49–68.
[47] R.W. Proctor, A. Dutta, Skill Acquisition and Human Performance, Sage Publications, Thousand Oaks, 1995.
[48] D. Gimmig, P. Huguet, J.P. Caverni, F. Cury, Psychonomic Bulletin and Review, in press.
[49] M.S. DeCaro, S.L. Beilock, Higher capacity can produce less optimal performance: Working memory, mental set, and strategy selection in math. Paper presented at the Annual Meeting of the Midwestern Psychological Association. Chicago, IL, 2006.
[50] A. Bogomolny, Modular arithmetic. Retrieved March 1, 2000 from http://www.cut-the-knot.com/blue/Modulo.shtml, 1996.
[51] A.R. Damasio, in: K.H. Pfenninger, V.R. Shubik (Eds.), The Origins of Creativity, Oxford University Press, Oxford, 2001, pp. 59–68.
[52] Y. Kareev, I. Lieberman, M. Lev, Journal of Experimental Psychology: General 126 (1997) 278–287.
[53] W. Gaissmaier, L.J. Schooler, J. Rieskamp, Journal of Experimental Psychology: Learning, Memory, and Cognition 32 (2006) 966–982.
[54] G. Wolford, S. Newman, M.B. Miller, G. Wig, Canadian Journal of Experimental Psychology 58 (2004) 221–228.
[55] A.S. Luchins, Psychological Monographs 54 (1942) 1–95.
[56] A.R.A. Conway, N. Cowan, M.F. Bunting, Psychonomic Bulletin and Review 8 (2001) 331–335.