Psychophysiology, 45 (2008), 119–129. Blackwell Publishing Inc. Printed in the USA.
Copyright © 2007 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2007.00605.x

Effects of incentive on working memory capacity: Behavioral and pupillometric data

RICHARD P. HEITZ,[a] JOSEF C. SCHROCK,[b] TABITHA W. PAYNE,[c] AND RANDALL W. ENGLE[a]

[a]School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, USA
[b]Division of Behavioral Sciences, Maryville College, Maryville, Tennesee, USA
[c]Department of Psychology, Kenyon College, Gambier, Ohio, USA

Abstract

We evaluated the hypothesis that individual differences in working memory capacity are explained by variation in mental effort, persons with low capacity exerting less effort than persons with high capacity. Groups previously rated high and low in working memory capacity performed the reading span task under three levels of incentive. The effort hypothesis holds that low span subjects exert less effort during task performance than do high spans. Subjects' pupil sizes were recorded online during task performance as a measure of mental effort. Both recall performance and pupil diameter were found to be increased under incentives, but were additive with span (incentives increased performance and pupil diameter equivalently for both span groups). Contrary to the effort hypothesis, task-evoked pupillary responses indicated that if anything, low span subjects exert more effort than do high spans.

Descriptors: Individual differences, Working memory capacity, Mental effort, Pupillometry

What is the basis of individual differences in working memory capacity? Despite decades of research, this question is still debated in the literature. With the emergence of working memory capacity as an explanatory mechanism in such diverse fields as clinical (e.g., Christopher & MacDonald, 2005), social (e.g., Schmader & Johns, 2003), developmental (e.g., Oberauer, 2005), as well as cognitive psychology, it is more important than ever to establish what working memory, as a system, represents. What is the basis for individual differences in working memory capacity, and why does it demonstrate reliable correlations to performance in other tasks? In the current work, we test a common, yet untested, explanation for individual differences in working memory capacity. Essentially, the argument is that some subjects are simply less willing to work hard, leading to lower performance in many tasks, working memory capacity measures included. Conversely, other subjects are willing to put forth the effort necessary to maintain high levels of performance. Hence, a spurious correlation arises: The hard workers contribute high working memory capacity scores as well as high levels of performance on criterion tasks.

There is reason to doubt such an explanation. Working memory capacity, although it does demonstrate correlations with a wide variety of tasks, is not promiscuous. Rather, correlations emerge in specific, predictable ways. In our own work, we have argued that working memory capacity will relate to task performance to the extent that the task requires high-level, controlled attention in the face of potential interference (Engle, 2002). Thus, we predict that a relationship will emerge in *any* task meeting that requirement. We have shown that working memory capacity is related to performance not only on memory tasks, but also lower level tasks that require the effortful control of attention in the face of interference (e.g., antisaccade, Eriksen flanker, and Stroop paradigms; see Unsworth, Schrock, & Engle, 2004; Heitz & Engle, 2007; Kane & Engle, 2003,respectively). In contrast, an effort hypothesis would make the prediction that working memory capacity should be related to performance in *all* tasks where effort is free to vary. We do not find such relationships in tasks that do not require controlled attention (e.g., the prosaccade trials on the *antisaccade task*, compatible trials in the Eriksen flanker task, and congruent Stroop trials). Unfortunately, these control conditions are also easier, leading to the possibility that some ceiling effect precluded the emergence of a correlation (but see Heitz & Engle, 2007).

To conclusively support an effort hypothesis of individual differences in working memory, one must be able to take a subject rated low in working memory capacity (a *low span* [LS]) and, with proper incentive, induce that person to perform like one rated high in working memory capacity (a *high span* [HS]). In other words, if high spans perform better than low spans because they are highly motivated, then providing additional motivation should abolish or attenuate this difference. However, if effort levels are *not* the key factor for bringing about individual

This research was supported by Grants F49620-97-1 and F49620-93-1-0336 from the Air Force Office of Scientific Research. We thank Nash Unsworth and Tom Redick for thoughtful comments and Michelle Grant for working with the participants.

Address reprint requests to: Richard P. Heitz, now at the Center for Integrative and Cognitive Neuroscience, Vanderbilt University, 301 Wilson hall, 111 21st Avenue South, Nashville, TN, 37203, USA. E-mail: richard.p.heitz@vanderbilt.edu or randall.engle@psych.gatech.edu

differences in working memory capacity, then additional incentives will have additive effects on performance. That is, both groups' performance may change as a result of increased incentive, but the difference between the groups will remain constant.

As well, one would hope to have an objective measure of effort expenditure independent of performance levels. This latter point is important, because failure to elicit changes in behavioral performance could indicate an ineffective or weak incentive manipulation. Also, such a measure would allow us to examine the hypothesis that despite the fact that LSs perform *worse* than HSs, they may be exerting *more* effort. This would certainly be predicted by any theory postulating some type of deficiency in low span subjects other than effort. Fortunately, such an objective measure exists in the pupillary response.

The Pupillary Response and Mental Effort

It is known that pupil dilation is sensitive to both within-task and between-task variation in effort. For example, Hess and Polt (1960, 1964) had subjects perform mental arithmetic on equations of increasing complexity. During a trial, pupil dilation was observed to follow a straightforward time course. The pupil began to dilate with the presentation of the problem, reached asymptote just before a response, then returned to baseline after the response. As the arithmetic equations became more difficult, the magnitude of the peak dilation at response increased. Extending this research, Kahneman and Beatty (1966;see also Ahern & Beatty, 1979; Kahneman & Peavler, 1969) presented participants with three tasks of increasing difficulty (digit recall of variable length, immediate word recall, and digit transformation). Again, pupillary dilations indexed both within-trial and between-task difficulty. In the digit recall task, which was of variable length, the pupil dilated with the presentation of each successive digit, reached asymptote just before recall, then constricted with the unloading of each digit. Between tasks, Kahneman and Beatty found that the difficult transformation task elicited the largest pupillary responses, followed by the short-term word task and the short-term digit task. Other researchers have confirmed these findings (e.g., Granholm, Asarnow, Sarkin, & Dykes, 1996; Peavler, 1974;for a review, see Goldwater, 1972) and shown that the pupil does not dilate during control tasks that hold processing load constant.

Previous research has attempted to address the relation of effort to cognitive performance using physiological measures. For instance, Ahern and Beatty (1979) reasoned that more intelligent individuals may simply be those people who always work harder; alternatively, they may have more automatized cognitive routines, freeing up resources for other simultaneous tasks. They used dilation of the pupil of the eye (explained in detail below) as a measure of effort expenditure. They found that low intelligence people exhibited larger pupil dilations, suggesting that this group actually worked harder than the high intelligence group. Other work has similarly addressed the effort–performance relationship using physiological measures other than pupil dilation. For instance, Larson, Saccuzzo, and Brown (1994) used monetary incentive to manipulate motivation during task performance while measuring such variables as heart rate and skin conductance. Although incentive seemed to have a small effect on ability measures, heart rate and skin conductance seemed unaffected. It seems likely that these measures are less sensitive than the pupillary response.

To test the effort hypothesis, we had high and low span subjects (previously measured by the Turner & Engle, 1989,operation span task) return to perform another working memory task (the reading span task, adapted from Daneman & Carpenter, 1980). During reading span, we manipulated incentive at three levels: no incentive, feedback, and feedback+monetary reward.[1] In the last condition, subjects earned bonus money based on their recall performance. At the same time, we recorded subjects' pupillary responses. Prior to the experiment, we took a *preexperimental* resting baseline. Additionally, we took a baseline recording at the beginning of each trial. These *trial baselines* were used to compute what is known as the *phasic* pupil response. Essentially, the trial baseline is subtracted from pupil size during the trial to yield task-evoked changes in pupil diameter. This response represents the change in effort, from baseline, during a given trial. So long as subjects exert more effort during a trial than during a resting interval, the phasic response will be positive. The use of such a difference score also eliminates global, or *tonic*, changes in pupil size unrelated to moment-by-moment changes in effort. This is not to say that tonic pupil size is unimportant; this measure is related to overall arousal levels and will prove to provide interesting data regarding span groups.

Again, the effort hypothesis states that, if given appropriate incentive, the difference between high and low span subjects in working memory capacity task performance should be eliminated or significantly attenuated. Furthermore, we would expect that in the standard case of no extra incentives, low spans exhibit smaller phasic pupil responses than high spans, indicating that the latter simply work harder. Alternatively, if effort is *not* the key factor leading to individual differences in working memory capacity, then the difference between high and low span performance should remain constant despite an effective incentive manipulation. Also, we might expect that low spans exhibit *larger* phasic responses than high spans during task performance, due to their deficiency in some underlying cognitive construct.

We would like to note outright that this study is not intended to support any particular theory of working memory, but rather, strives only to test a mental effort hypothesis. Thus, this study is theoretically neutral, so long as our two measures (operation span; Turner & Engle, 1989; and reading-span; Daneman & Carpenter, 1980) are assumed to be adequate measures of working memory capacity.

Method

Participants

Participants were recruited from the surrounding Atlanta area through newspaper advertisements or from the Georgia Tech undergraduate subject pool. Subjects were never tested twice in one day, and the interval between Session 1 and Session 2 ranged from 1 day to several months. All individuals were native English speakers, had corrected-to-normal vision, and were not taking any medication known to affect memory or attentional focus. As well, subjects were screened so as to be free of psychiatric and

[1] The three incentive conditions, no feedback, feedback, and feedback+monetary incentive, were actually conducted as three separate experiments. However, because we treat experiment as a factor to test the effects of incentive, they have been combined here into a single model. The experiments were similar enough that treating them as different conditions does not present a problem.

Table 1. *Demographic Means (Standard Deviations) for Each Incentive Condition and Working Memory Span Group*

	Working memory span group	
Incentive condition	Low span	High span
No feedback		
Ospan	6.33 (2.60)	24.52 (5.88)
Age	24.37 (4.74)	22.10 (3.87)
% Female	56.7%	40.0%
N	30	30
Feedback		
Ospan	6.07 (2.12)	24.00 (5.18)
Age	26.60 (5.56)	23.53 (5.23)
% Female	63.3%	43.3%
N	30	30
Feedback+monetary incentive		
Ospan	6.32 (2.17)	24.04 (4.62)
Age	23.40 (4.74)	21.24 (2.91)
% Female	36%	60%
N	25	25

neurological dysfunction[2] and were rated as having 20/20 or better vision by Snellen chart. Subjects were between the ages of 18 and 35 (Table 1).

Stimuli and Procedures

Participants performed the operation span task (OSpan; Turner & Engle, 1989) to assess working memory capacity. The OSpan task has been shown to have good internal consistency and test–retest reliability (Conway et al., 2005; Engle, Tuholski, Laughlin, & Conway, 1999; Klein & Fiss, 1999). Those falling in the upper and lower quartiles of this distribution were designated high and low spans, respectively. The OSpan task is part of an ongoing screening procedure; the quartiles computed for LS/HS cutoffs include an *N* greater than 3000. From this large pool of individuals, 85 high spans (OSpan $M = 24.2$, $SD = 5.2$) and 85 low spans (OSpan $M = 6.2$, $SD = 2.3$) were identified and asked to return on a later date to perform a modified version of the reading span task (Daneman & Carpenter, 1980). Subjects were assigned to incentive conditions as follows: no incentive (30 HS, 30 LS), feedback (30 HS, 30 LS), feedback+monetary incentive (25 HS, 25 LS).

Operation-span task. The stimuli were identical to those used by Engle et al. (1999). Participants viewed strings of simple arithmetic problems that were each followed by a single, high-frequency word. The operations were of moderate difficulty (Conway & Engle, 1996). Participants first read the equation out loud, responded *yes* or *no* as to whether the equation was true or false, and then read the word out loud. For example, a string might appear as *IS (4/2)+3 = 5 ? BIRD*. One would respond, *Is four divided by two plus three equal to five? Yes. Bird.* The experimenter then pushed a key and the next operation–word pair appeared. After a set of from two to five operation–word pairs, a series of question marks appeared as a cue for recall. Participants were instructed to recall the words, by writing them on an answer sheet, in the same order they were presented.

The set sizes were initially randomized, and all subjects received the same order. Each set size was presented three times for a total of 12 trials. Three practice trials of set size two preceded the experimental trials. An individual's span score was calculated as the sum of all *perfectly* recalled set sizes. So for example, if an individual recalled perfectly all trials at set size two, and two out of three words on all of the trials at set size three, the span score would be 6 (2+2+2+0+0+0). In addition, all individuals were required to maintain an accuracy of 85% for the operation strings. Failure to meet this criterion excluded the participant from all further experiments and data analyses.

Reading-span task. Subjects classified as high or low by the OSpan task were asked to return to perform the reading-span task. Table 1 shows means and standard deviations of OSpan scores for each of the three incentive conditions. Prior to beginning the experiment, subjects were dark adapted for approximately 5 min. Before reading task instructions (but after completing informed consent), participants were calibrated on the eye-tracking equipment. As part of this calibration, subjects fixated on a+sign for 7 s. This provided a measure of each individual's preexperimental, baseline pupil diameter, also known as tonic pupil size. This was our measure of resting arousal levels not affected by incentive condition or cognitive activity (subjects sat passively).

The reading-span task consists of sentences subjects must read aloud, each followed by a single letter (also read aloud). To facilitate pupillometric recordings, we partitioned the reading-span task into a series of segments, with each trial segment presented on a separate screen. The first segment was a 7-s baseline consisting of a+followed by a series of ~ symbols. The length of the string approximated the length and luminance[3] of the following sentence. Participants were asked to focus on the+sign until the first sentence appeared; this ensured that subjects did not need to saccade to begin the trial. This period provided a baseline pupil size for each individual trial. Subtracting this baseline from pupil size during the trial provided a measure of mental effort expenditure (the phasic response). Following the baseline period, an individual sentence appeared, and subjects were to read this sentence aloud. The sentences were grammatically simple (Flesh-Kincaid grade level = 4.7; e.g., *Jim's mother finally agreed to let him have a dog as a pet*) and remained visible until oral recitation was completed. On the next screen, subjects observed a single letter, sampled (without replacement within a trial) from a pool of 12 letters: {F, P, Q, J, H, K, T, S, N, R, Y, L}. Subjects encountered sets of two to seven sentence–letter pairs. The recall segment was cued by a set of three question marks (???); subjects were to verbally recall the letters in the same order they were presented. In addition, they were instructed to say *blank* for any letter they did not remember and *done* when finished recalling. The experimenter keyed each letter as participants recalled them, and letters remained visible until the experimenter pressed the *enter* key (which occurred when the participant said *done*). Prior to this, participants were able to correct their responses.

The trial structure somewhat depended on incentive condition, although the first half of each trial was identical: a baseline period (7000 ms), sentence(*n*) (duration subject controlled),

[2] Other demographic characteristics, such as smoking status and anxiety levels, were not recorded. Although one could make the case that these variables are systematic with span (e.g., low span subjects may be more anxious or more likely to self-medicate through nicotine), the prediction would be *larger* tonic pupil size for low span subjects. The data will be in opposition to this.

[3] Although luminance values were not strictly held constant throughout segments of the trial, there is no reason to think that this will introduce any working memory span-related confounds. All subjects viewed identical stimuli presented on the same computer monitor.

letter(*n*) (2500 ms), and recall (duration subject controlled). For instance, a set size of two would proceed as baseline, sentence1, letter1, sentence2, letter2, recall. In the *no feedback* condition, recall was followed immediately by a comprehension question, based randomly on one of the previously viewed sentences. Subjects were to respond *yes* or *no*, after which subjects viewed comprehension question feedback for 2000 ms. In the *feedback* condition, recall was followed immediately with *You recalled* × *out of* y *letters correctly*, which remained visible for 1500 ms. Then, subjects performed the comprehension question and received comprehension question feedback for 2000 ms. The *feedback+money* condition was identical to the feedback condition except for the addition of earnings information. During recall feedback, subjects were also told how much the trial was worth based on their performance, as well as their cumulative earnings thus far. This was visible for 4000 ms. Then, subjects performed the comprehension question. The comprehension feedback screen also included what subjects had earned for that trial. This remained on-screen for 3000 ms. There was no intertrial interval; the next trial began immediately upon offset of the comprehension screen feedback for all incentive conditions.

In the feedback+money condition, participants were paid based on both their letter recall performance and whether or not the comprehension question was answered correctly. For letter recall, an incremental payment procedure was employed (see Table 2). The value of any trial was expressed by

$$\sum_{i=1}^{C} (.06 \times i)$$

where *C* is the number of correctly recalled letters for that trial. For instance, the value of the first letter correct was 6 cents, the second 12 cents, the third 18, and so on. The total value for a set in which four letters were recalled correctly would be (.06+.12+.18+.24), or 60 cents. Thus, while the increment was held constant between letters, the trial value was compounded. As illustrated in Table 2, a single trial could be worth up to $1.68. Over the course of the experiment (24 trials), this amounted to a possible $20.00 bonus in addition to the $20.00 (or course credit) all individuals received for participation (earnings were rounded up to the nearest $5.00 increment). Following recall, participants were shown how much bonus money the trial was worth, given their performance. This was termed *total possible for trial*. If participants answered the comprehension question correctly, the total possible was added to a running total, also displayed on-screen for 4000 ms. If incorrect, participants earned only half of the total possible. Thus, to maximize bonus money, subjects had to remember all the letters *and* understand all the sentences. Note also that individuals could not predict any trial set size and so did not know how much any trial was worth until the recall phase. The running total, total possible, and trial earnings were displayed on-screen only after the recall phase and disappeared at the start of the next trial. At the conclusion of the experiment, participants' total bonus earnings were rounded up to the nearest $5.00, and they were paid. Accordingly, if an individual earned $5.25, he or she would receive $10.00.

Table 2. *Trial Earnings as a Function of Number of Letters Recalled Correctly and Comprehension Question Accuracy*

	Comprehension question accuracy	
Letters correct	Correct	Incorrect
1	$0.06	$0.03
2	$0.18	$0.09
3	$0.36	$0.18
4	$0.60	$0.30
5	$0.90	$0.45
6	$1.26	$0.63
7	$1.68	$0.84

It is noteworthy that although the next trial began immediately after feedback (i.e., there was no intertrial interval), the latency between the end of the trial and the next baseline period differed depending on incentive condition. Subjects in the no feedback and feedback conditions encountered only the 2000-ms interval required to present comprehension question accuracy (*correct* vs. *incorrect*), after which the next baseline immediately began. In the feedback+money condition, however, subjects were additionally allowed time to view their trial earnings. An extra 1000 ms (3000 ms total) was allotted for this.

The set sizes were initially randomized, and each subject received the same order. Each set size was presented four times for a total of 24 experimental trials. Three practice trials of set size 2 were presented prior to the experimental block. A subject's span score was the total number of letters recalled in the correct order.

Stimuli were presented in white against a black background on a 17-in. monitor. Participants were tested individually with an experimenter present and sat approximately 36 in. from the screen. To minimize the effects of extraneous sounds, a white noise was presented during the entire experimental session. This was presented by white-noise-generating equipment located directly below the computer monitor. As well, ambient light was limited to a single small desk lamp located well behind the participant. Luminance values were not recorded for ambient light or for any stimuli presented in this experiment. Although this will affect pupil sizes, it does not present a problem for group comparisons, as all subjects viewed identical stimuli presented on the same computer monitor. A program written in E-Prime version 1.0 presented all stimuli, recorded accuracy data, and controlled the eye-tracking unit. The entire session typically required about 45 min, and rarely exceeded 1 h.

Eye-tracking equipment. All participants performing the reading-span task were tracked using an Applied Science Laboratories model 5000 eye-tracker unit sampling at 60 Hz. A magnetic head-tracker controlled for head movement. Pupil data were recorded using proprietary software provided by ASL. Eye-fixation data were recorded but will not be reported. Pupil data were filtered for blinks and any momentary loss of calibration. No subjects or trials were eliminated due to excessive blinking. From this raw data set, the following submeasures were extracted: tonic pupil size prior to beginning the experiment proper (the preexperimental baseline), tonic pupil size recorded at the beginning of each of the 24 trials (trial baselines), and phasic pupillary response during sentence reading, during letter encoding, during the first 6 s of recall, and during feedback (for the feedback and feedback+money conditions only). Phasic responses were computed as difference scores by subtracting the mean baseline pupil size for that trial from the maximum pupil size during a processing epoch (Beatty & Lucero-Wagoner, 2000; Verney, Granholm, & Dionisio, 2001). Thus, phasic responses were corrected for changes in tonic pupil size throughout the experiment.

This experiment was thus a 2 (working memory span: high vs. low) × 3 (incentive: no incentive, feedback, feedback+money) × 6 (set size: 2–7) mixed design with span and incentive condition as between-groups variables and set size as a within-subject variable. The dependent measures were recall accuracy and pupil dilation from baseline. Several submeasures were extracted from the pupil data: baselines (one for each trial = 24 total, as well as 1 preexperimental baseline), dilation during sentence reading (one to seven sentences), during letter encoding (one to seven letters), and during recall (set sizes of from two to seven). Again, the preexperimental baseline was recorded after informed consent but before the task, and trial baselines were recorded prior to each trial. Statistical analyses employed analysis of variance (ANOVA) and, where mentioned, analysis of covariance (ANCOVA). All repeated-measures analyses employed the Greenhouse–Geisser correction for violation of sphericity.

Results

Behavioral Data

Operation span and reading span were correlated, $r = .61$, $p < .001$. This is consistent with the view that the tasks measure a common construct. The validity of this argument has been established in a number of papers (e.g., Engle, Cantor, & Carullo, 1992; Engle et al., 1999; Turner & Engle, 1989).

Recall performance. Figure 1 presents recall performance as a function of set size, span group, and incentive condition. As is evident, high spans recalled significantly more letters than did low spans, $F(1,164) = 137.67$, $p < .001$, partial $\eta^2 = .46$, and the incentive conditions significantly increased performance, $F(2,164) = 10.42$, $p < .001$, partial $\eta^2 = .11$. Contrasts revealed that performance for the feedback condition was marginally better than for the no feedback condition, $F(1,164) = 3.139$, $p < .08$, partial $\eta^2 = .02$, and the feedback+monetary reward condition was significantly better than the feedback condition, $F(1,164) = 20.64$, $p < .001$, partial $\eta^2 = .111$. However, these incentive-related increases in performance did *not* interact with span, $F(2,164) < 1$, n.s., partial $\eta^2 = .01$, nor was there an interaction between span, incentive condition, and memory load, $F(10,820) < 1.0$, n.s., partial $\eta^2 = .003$. The mean difference between high and low spans were, for no feedback, $M_{\text{diff}} = 1.11$, for feedback alone, $M_{diff} = 1.17$, and for feedback+money, $M_{diff} = .94$. Hence, although incentive did increase performance, the difference between highs and lows remained constant. And it is clear that low span performance under the highest level of incentive was below the performance level of high spans under no incentive. Thus, although incentives do have an impact on performance levels, they do not attenuate span differences.

Sentence viewing times. Presented in Figure 2 are average sentence viewing times throughout trials. Recall that sentence viewing time was subject controlled; each sentence remained visible until the subject completed reading it out loud. Memory load, plotted on the x-axis, was 0 for the first sentence, as no letter had yet been presented. Not only was there a main effect of memory load, $F(6,984) = 98.55$, $p < .001$, partial $\eta^2 = .38$, but a linear trend indicated that viewing times tended to increase with memory load, $F(1,164) = 162.03$, $p < .001$, partial $\eta^2 = .50$. Evident also was a main effect of span, $F(1,164) = 13.16$, $p < .001$, partial $\eta^2 = .07$, indicating that, overall, low spans tended to require more time to read the sentences. Averaging across incentive condition, low spans required 5.4 (0.28) s, whereas high spans required 5.0 (0.37) s. A marginal main effect of incentive condition also emerged, $F(2,164) < .09$, partial $\eta^2 = .03$. Contrasts revealed that, although the no feedback and feedback alone conditions were statistically equivalent, viewing times were significantly slower in the feedback+money condition, $F(2,164) = 2.50$, $p < .05$, partial $\eta^2 = .03$. Additionally, a Memory Load × Incentive Condition interaction emerged, $F(12,984) = 1.99$, $p < .05$, partial $\eta^2 = .02$. Contrasts revealed that for both span groups, viewing times increased more quickly in the feedback+money condition relative to the other two conditions, $F(2,164) = 3.29$, $p <.\ 05$, partial $\eta^2 = .04$. There was, however, no Span × Memory Load interaction, suggesting that viewing times for both groups increased at approximately the same rate, $F(6,984) = 1.20$, n.s., partial $\eta^2 = .01$. The three-way interaction between span, incentive condition, and memory load did not attain significance, $F(12,984) < 1.0$, n.s., partial $\eta^2 = .01$.

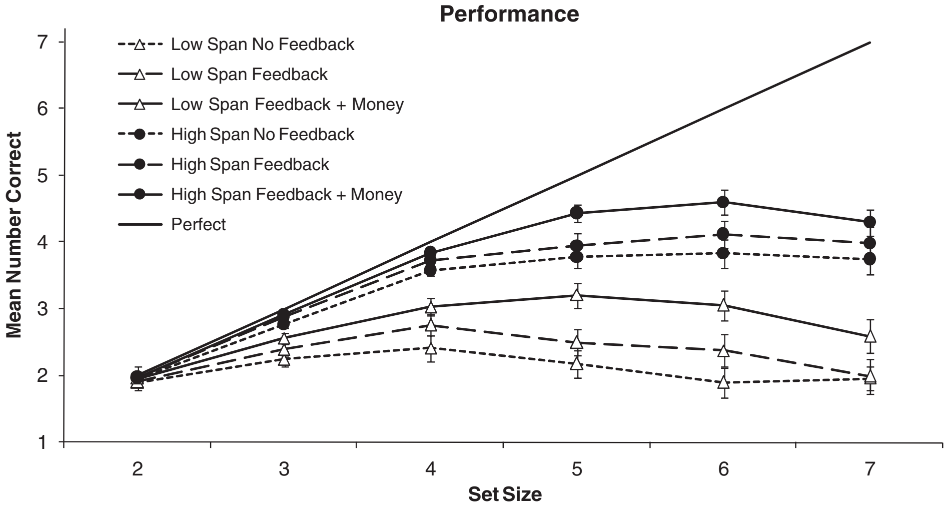

Figure 1. Recall performance functions for high and low span individuals across the three levels of incentive. Vertical bars represent ± 1 standard error of the mean.

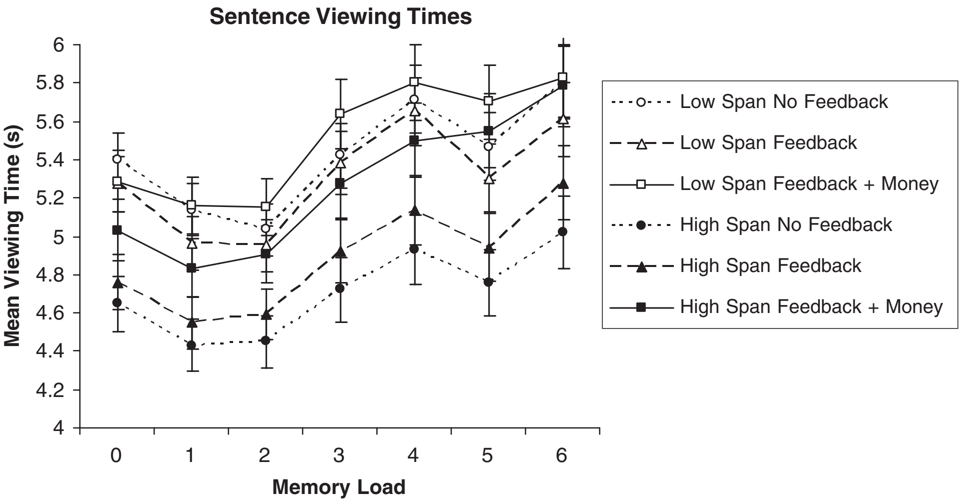

Figure 2. Sentence viewing times. Vertical bars represent ± 1 standard error of the mean.

Comprehension question accuracy. Following recall, subjects were presented with a single comprehension question regarding one of the sentences from that trial. As might be expected, comprehension question accuracy tended to decline as memory load increased, $F(5,820) = 28.80$, $p < .001$, partial $\eta^2 = .15$. Although accuracy rates were quite high, high spans exhibited a slight advantage (high span mean proportion correct = .92; low span mean proportion correct = .89), $F(1,164) = 9.88$, $p < .01$, partial $\eta^2 = .06$. These proportions were equivalent to approximately 2.64 incorrect responses for low spans and 1.92 incorrect responses for high spans. There was no apparent Span × Incentive interaction, $F(2,164) = 2.38$, n.s., partial $\eta^2 = .03$ nor did span group interact with memory load, $F(5,820) = 1.55$, n.s., partial $\eta^2 = .01$. Memory load also did not interact with incentive condition, suggesting that although error rates increased with memory load, it was statistically equivalent for each condition, $F(10,820) < 1.0$, n.s., partial $\eta^2 = .01$. The three-way interaction between span group, memory load, and incentive condition also did not attain significance, $F(10,820) < 1.0$, n.s., partial $\eta^2 = .01$.

Pupil Data

Preexperimental baselines. Figure 3 presents preexperimental baselines for high and low span participants, separately for each experimental condition. Three subjects were omitted from these analyses due to technical difficulties (they did not supply a preexperimental baseline). Mean pupil diameter was roughly equivalent across the three incentive conditions; as this measure was recorded prior to any experimental manipulation, this is not surprising, $F(2,164) = 1.32$, n.s., partial $\eta^2 = .02$. What is remarkable, however, is the difference in tonic pupil size between high and low span subjects, $F(1,164) = 9.52$, $p < .01$, partial $\eta^2 = .06$. Again, this measure was recorded prior to beginning the experiment; subjects simply passively viewed a + symbol. It is known that tonic pupil diameter increases from the age of 20 onward at a rate of approximately 0.4 mm/decade (Bourne, Smith, & Smith, 1979). The mean age difference between high and low span subjects was only 2.5 years. Nevertheless, we examined the relationship between Ospan score, baseline pupil

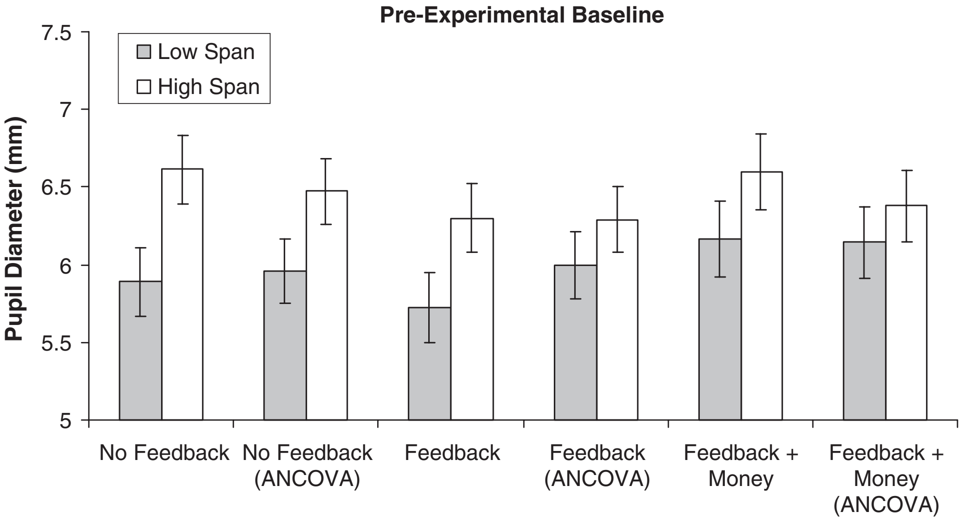

Figure 3. Preexperimental baseline pupil diameter across the three incentive conditions before and after an ANCOVA holding age constant. Vertical bars represent ± 1 standard error of the mean.

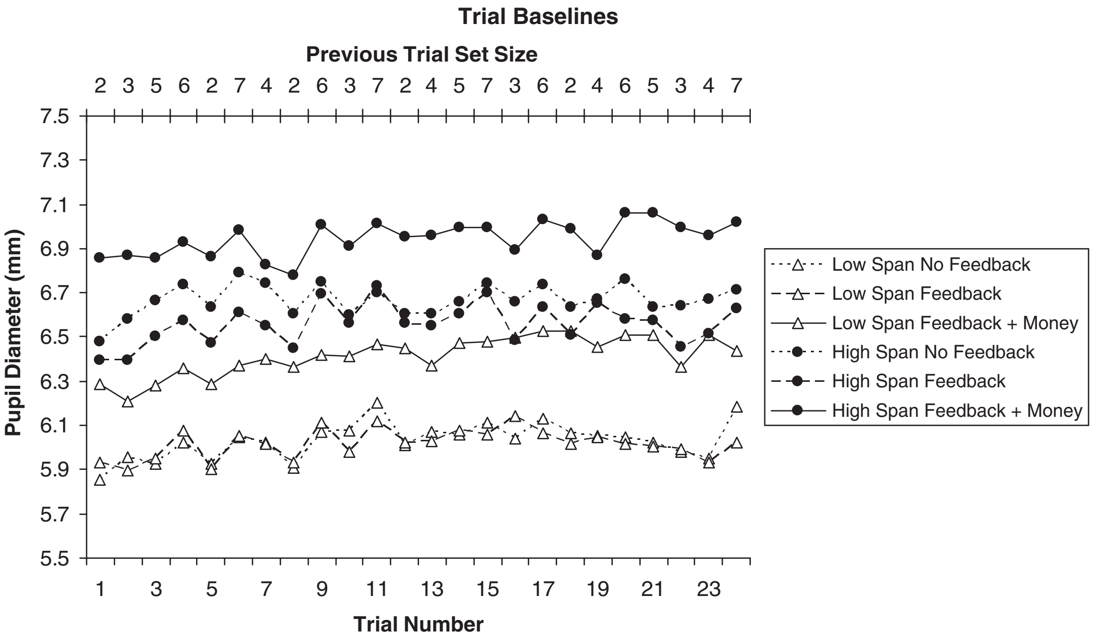

Figure 4. Baseline pupil diameter for each of the 24 trials. The top horizontal axis presents the set size for the previous trial. Error bars have been withheld for display purposes.

size, and age. Though small, there was a significant correlation between Ospan score and age, $r(167) = -.249$ and between Ospan and pupil diameter, $r(167) = .237$. However, the strongest relationship emerged for pupil diameter and age, $r(168) = -.407$. To control for the possibility that the group differences in baseline pupil size were due to age, we computed the Ospan–pupil diameter partial correlation holding age constant. This resulted in a decrease from $r(167) = .237$, $p < .01$ to $r(164) = .154$, $p < .05$. This suggests that, although some of the group differences in tonic pupil size are due to age, there exists important residual variance captured by Ospan score. Figure 3 presents the residual differences in tonic pupil size after controlling for age (using ANCOVA).[4] Entering age as a covariate yielded a marginal main effect of span group, $F(1,163) = 3.62$, $p < .06$, partial $\eta^2 = .02$. There was no Span × Incentive Condition interaction, $F(2,163) < 1.0$, n.s., partial $\eta^2 = .003$.

Trial baselines. As described above, subjects began each trial by focusing on a fixation screen for 7 s. The fixation screen appeared as a+sign followed by ~ symbols, approximating the length and luminance of the following sentence. Figure 4 presents these mean values as a function of trial number, span group, incentive condition, and previous trial set size (upper x-axis). It is apparent that the group differences observed in the preexperimental baselines persist throughout the experiment; a main effect of span was observed, $F(1,153) = 7.99$, $p < .01$, partial $\eta^2 = .07$.

Additionally, it is evident in Figure 4 that the trial baselines were much larger for the feedback+money condition than either the no feedback or feedback conditions, $F(2,154) = 2.54$, $p < .05$, partial $\eta^2 = .03$. This would be predicted on the basis that monetary incentive increases effort levels and is certainly consistent with the performance data presented above. A follow-up ANCOVA, again holding age constant, drove the effect of incentive to marginal significance ($p = .09$), although upon inspection, the mean pupil diameter in the feedback+money condition was still quite large relative to the average of the other two conditions (low span: $M_{NoFeed\&\mathrm{Feed}} = 6.02$ mm, $M_{\mathrm{Feed+Money}} = 6.42$ mm; high span: $M_{\mathrm{NoFeed\&Feed}} = 6.61$ mm, $M_{\mathrm{Feed+Money}} = 6.95$ mm). There was no apparent interaction with span, which further supports the conclusion that incentive has equivalent effects on both span groups, $F(2,154) < 1.0$, n.s., partial $\eta^2 = .00$.

As Figure 4 also illustrates, there is an intriguing degree of similarity in the peaks and troughs of the functions. The top x-axis of Figure 4 lists the set size on the previous trial. Clearly, the functions drop as the previous trial set sizes decrease and rise as previous set sizes increase. Previous literature assumed that during an intertrial interval, pupil dilation quickly returns to baseline (e.g., Beatty, 1986). However, the present data suggest that pupil size is affected on a more long-term basis by the amount of effort required by the previous trial. Table 3 presents correlations between previous trial set size and trial baseline pupil diameter as a function of span and incentive condition. These correlations were run on the corrected values from ANCOVA, holding age constant. (Correlations run on uncorrected values were equivalent, and in no case did any conclusions differ). As well, note that each correlation has 22 degrees of freedom, having been collapsed across intersubject variability. Although this has the effect of

Table 3. *Correlations between Trial Baseline Pupil Diameter and Previous Trial Set Size*

	Incentive condition		
Span group	No feedback	Feedback	Feedback+monetary incentive
High span	.78*	.78*	.55*
Low span	.68*	.59*	.29

*$p < .01$

[4] All pupillary analyses and figures to be reported below use ANCOVA holding age constant. The reader is assured that our conclusions would be identical had we not controlled for age.

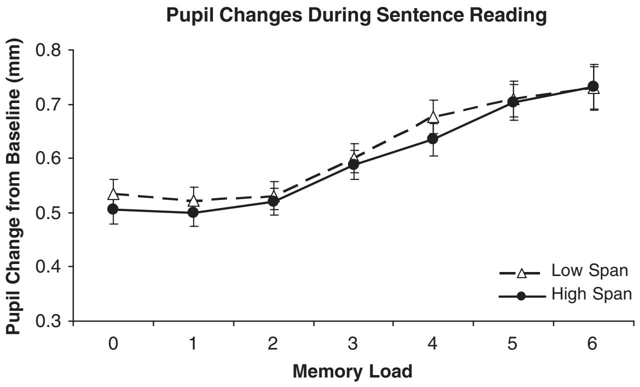

Figure 5. Pupil change from baseline during sentence reading. Vertical bars represent ± 1 standard error of the mean. These data have been collapsed across incentive condition. Note that the first memory load is 0, as the first sentence occurs before presentation of the first to-be-recalled letter. As there was no interaction with incentive condition, these data have been collapsed.

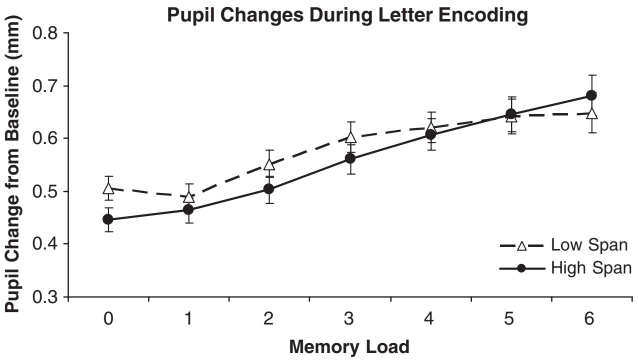

Figure 6. Pupil change from baseline during letter encoding. Vertical bars represent ± 1 standard error of the mean. These data have been collapsed across incentive condition. Note that the first memory load is 0, as the first letter occurs before there is any memory load.

somewhat inflating the correlations, it has the advantage of filtering out noise inherent in the biological signal. As can be seen, the peaks and troughs in Figure 4 are not random fluctuations but reflect changes in pupil size due to the difficulty of the *previous* trial. We will argue later that this reflects significant carryover effects in the phasic response. In other words, the pupil does not quickly return to resting levels.

Phasic responses during sentence reading. Figure 5 presents mean phasic pupil responses during sentence reading as a function of memory load. All 24 trials contributed to memory loads of 0 and 1, as subjects were always presented with a minimum of 2 sentence/letter pairs. Conversely, only four trials were presented at a set size of 7; thus, there are fewer observations at larger memory loads. Note that the final sentence was associated with a memory load of 6. A 2 (span) × 7 (memory load) ANCOVA (holding age constant) yielded a main effect of memory load, $F(6,978) = 6.98$, $p < .001$, partial $\eta^2 = .04$, as well as a significant linear trend, $F(1,163) = 10.93$, $p < .01$, partial $\eta^2 = .06$, indicating that pupil diameter tended to increase with memory load. There was no significant main effect of span, $F(1,163) < 1.0$, n.s., partial $\eta^2 = .001$, nor of incentive condition, $F(2,163) < 1.0$, n.s., partial $\eta^2 = .01$. As well, incentive condition did not interact with span group, $F(2,163) = 1.57$, n.s., partial $\eta^2 = .02$.

Memory load, although having an effect on its own (mentioned above), did not interact with span group, $F(6,978) < 1.0$, n.s., partial $\eta^2 = .003$, nor with incentive condition, $F(12,978) = 1.34$, n.s., partial $\eta^2 = .02$. Similarly, no Span × Incentive × Memory Load interaction appeared, $F(12,978) < 1.0$, n.s., partial $\eta^2 = .01$.

Phasic responses during letter encoding. Figure 6 presents mean phasic pupil responses during letter encoding. This reflects the increase in pupil size as subjects view letters within a trial. A 2 (span) × 7 (memory load) ANCOVA indicated a main effect of memory load, $F(6,978) = 7.23$, $p < .001$, partial $\eta^2 = .04$, and a significant linear component, $F(1,163) = 12.92$, $p < .001$, partial $\eta^2 = .07$. Again, this supports the observation that pupil sizes tended to increase with memory load. As is suggested by Figure 6, low spans exhibited larger phasic responses than high spans at memory load of 0 through 3. It also appears that this effect slightly reverses at the largest set size, giving rise to a Span × Memory Load interaction, $F(6,978) = 2.89$, $p < .05$, partial $\eta^2 = .02$. Follow-up tests, however, could only confirm a significant difference at a memory load of 0, $t(168) = 3.33$, $p < .001$. There was no main effect of span, $F(1,163) < 1.0$, n.s., partial $\eta^2 = .002$, incentive condition, $F(2,163) < 1.0$, n.s., partial $\eta^2 = .01$, nor a Span × Incentive Condition interaction, $F(2,163) = 1.66$, n.s., partial $\eta^2 = .02$.

Although memory load was observed to interact with span (described above), it did not interact with incentive, $F(12,978) = 1.28$, n.s., partial $\eta^2 = .02$, nor was there a Span × Incentive × Memory Load interaction, $F(12,978) = 1.01$, n.s., partial $\eta^2 = .01$.

Phasic responses during recall. The duration of the recall period was unconstrained; subjects could take as long as they wished to recall the letters and were also allowed to change their answers prior to ending recall (see Methods). For this reason, we considered only the first 6 s of the recall period. It was within this interval that subjects were at their maximal load for the trial and began verbal recall. Figure 7 presents these data. Note that memory load, on the *x*-axis, begins at 2 and continues to 7. Unlike the pupillary data for sentence reading and letter encoding, the recall period occurs only once per trial. The lowest possible value for this would be a set size of 2. (In contrast, pupil data for, say, sentence reading depicted changes *during* a trial). ANCOVA revealed a significant main effect of memory load, $F(5,815) = 3.11$, $p < .\ 05$, partial $\eta^2 = .02$, with a significant linear trend, $F(1,163) = 6.68$, $p < .05$, partial $\eta^2 = .04$. Once again, this indicates that pupil diameter increased as set sizes increased. Also evident was a Span × Incentive × Memory Load interaction, $F(10,815) = 2.08$, $p < .05$, partial $\eta^2 = .03$. Figure 7a indicates that under no incentive, low span subjects exhibit larger phasic pupillary responses than do high spans, $F(1,57) = 7.46$, $p < .01$, partial $\eta^2 = .12$, though there was no Span × Memory Load interaction, $F(5,285) = 1.11$, n.s., partial $\eta^2 = .02$. Figure 7b depicts data for the feedback condition. Here, there is no main effect span, though a Span × Memory Load interaction did emerge, $F(5,285) = 2.42$, $p < .05$, partial $\eta^2 = .04$. It appears from Figure 7b that phasic responses are larger for low spans during smaller set sizes. Indeed, low spans had significantly larger pupil sizes at set sizes 2, $t(58) = 2.52$, $p < .05$, and 3, $t(58) = 2.98$, $p < .05$. Finally, Figure 7c presents pupillary responses for the feedback+money condition. Again, there was no main effect of span but a significant Span × Memory Load interaction, $F(5,235) = 2.77$, $p < .05$, partial $\eta^2 = .\ 06$. This interaction ap-

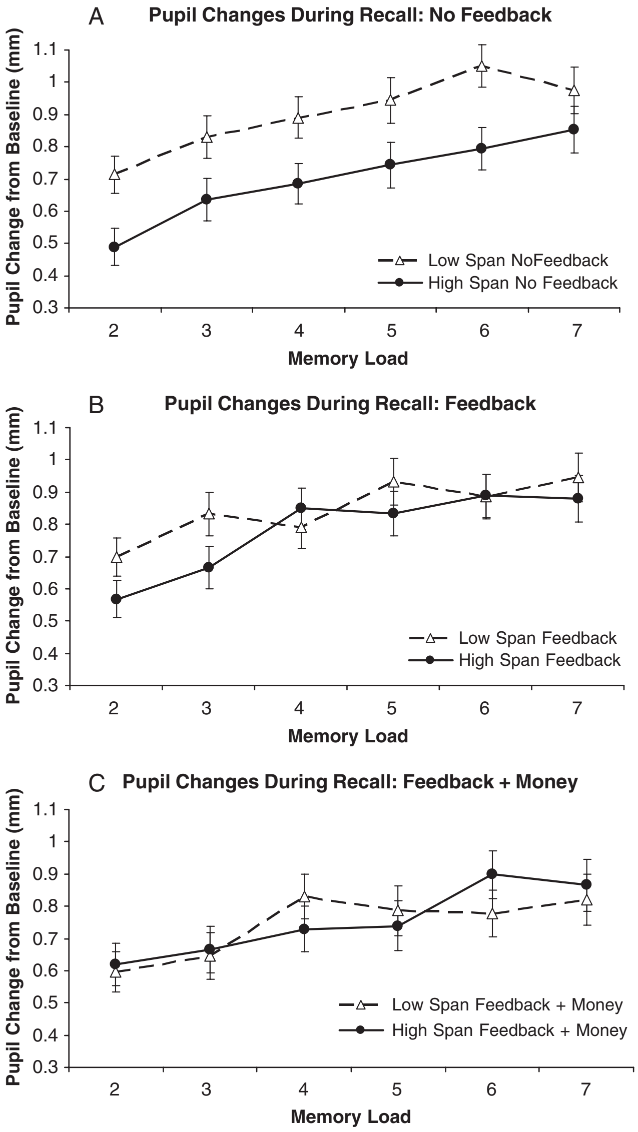

Figure 7. Pupil change from baseline during recall. Vertical bars represent ± 1 standard error of the mean. a: No feedback; b: feedback; c: feedback+money.

pears to be due to larger phasic responses for low spans at a memory load of 4, $t(48) = 1.94$, $p < .06$, and smaller phasic responses for low spans at a memory load of 6, $t(48) = -2.60$, $p < .05$.

Discussion

The purpose of the present study was to evaluate whether or not, and to what extent, high and low span subjects differ in the amount of effort expended during working memory task performance. Disregarding working memory span groups, we found that incentive did have strong effects on performance in the reading span task. Performance in the feedback+monetary reward condition was significantly higher than either the no feedback or feedback alone conditions. Importantly, the extent of increase was statistically identical for high and low spans, and mean comparisons revealed that the difference between the two groups remained constant across incentive levels. Hence, high span subjects are not simply those who are more motivated than low spans; had this been the case, then the difference between the two groups should have lessened under incentive. Of course, this conclusion is only valid to the extent that one can show that incentives did, in fact, increase effort. We used dilation of the pupil of the eye as our measure of effort expenditure. Research shows that the pupil is sensitive to changes in mental effort, being higher in tasks or conditions requiring more resources and constant in tasks holding processing load constant.

The pupil data presented here support two primary conclusions. First, tonic pupil size was sensitive to changes in incentive. Baseline measurements recorded at the beginning of each trial (trial baselines) were larger for the feedback+money condition than either the feedback or no incentive conditions. This is important, as it suggests that the increase in behavioral performance was related to an increase in arousal levels, as indicated by tonic pupil diameter. That said, the difference between high and low span subjects in both behavioral performance and pupil diameter was equivalent across all three incentive conditions. Stated differently, incentive had equally arousing effects in high and low span subjects, and this arousal had equal effects on behavioral performance.

Second, and most to the point, the phasic pupillary response indicated that high spans are *not* simply those individuals who expend more effort during task performance. Had this been true, high spans should have shown a larger increase in pupil diameter (relative to that trial's baseline) than did low spans. If anything, the *reverse* holds true. For instance, during letter encoding, low spans exhibited a larger phasic response during smaller set sizes, but lower phasic responses during the largest set size. Though speculative, one might take this as evidence that low spans must work harder than high spans to reach a level of performance that is inferior to high spans. As well, one might speculate that low spans begin to give up at larger set sizes, giving rise to the downward trend in those pupillary data. The more important data concern pupillary responses during recall. As mentioned, we feel that this is the point of maximal load and so reflects the best estimate of effort expenditure during the task. Unlike the pupillary responses during sentence reading and letter encoding, the phasic response did interact with incentive condition during recall. In the standard condition of no incentives, there exists a rather striking difference—low span subjects exhibit much larger phasic responses than do high spans. We take this to mean that low spans actually work *harder* than high spans—a result clearly in opposition to the mental effort hypothesis. It is also clear that adding any type of incentive diminishes this difference. In the feedback and feedback+money conditions, phasic responses for high and low span subjects were much more similar. A close look at Figure 7b suggests that, like the phasic response during letter encoding, low span subjects may exert more effort during smaller set sizes. This effect is not apparent in the feedback+money condition (Figure 7c), though it cannot be ruled out that low spans produce sometimes larger and sometimes smaller phasic responses, relative to high spans.

These data are clearly in opposition to the mental effort hypothesis. Incentive affects performance equally for high and low span groups, as indicated by both behavioral performance and pupil diameter. Phasic pupillary responses recorded at different segments of a trial indicated that, if anything, *low spans* exert more effort.

Although we can now rule out the hypothesis that high spans perform better than low spans simply due to more effort expenditure, we cannot discount the possibility that high spans have, in general, higher arousal levels than low spans. Our data revealed that high spans have larger resting pupil size than do low spans. This was true both for the baselines recorded at the beginning of each trial as well as for the preexperimental baseline. The preexperimental baseline is important because subjects had only limited information about the nature of the experiment and no incentive manipulations had yet been employed. Furthermore, subjects sat passively and viewed a+symbol for 7 s; in other words, subjects had no real cognitive task. Partial correlation analysis revealed that part of this relationship is due to age. However, after controlling for age there remained a residual correlation between OSpan score and tonic pupil size. What significance does this baseline difference have for our interpretation of high span/low span differences? Some have argued that tonic pupil size is a measure of global arousal levels (Granholm & Steinhauer, 2004) whereas early work on attention and effort (Kahneman, 1973) suggested that arousal levels are somehow related to capacity. This would be quite consistent with a working memory *capacity* viewpoint. According to our view, high spans are those individuals who have a greater ability (relative to low spans) to control attention in interference-rich situations. Hence, we do not view capacity as an amount per se, but rather the extent to which an individual can willfully control attention. It may be the case that arousal levels are somehow correlated with this ability. However, to say that high spans outperform low spans simply because they are more aroused seems to us to be an oversimplification. The situation is likely more complicated. Arousal levels, as measured by tonic pupil size, are affected by many factors, capacity being just one of a host of possible constructs.

The Persistence of the Phasic Response

We observed that baseline pupil size at the beginning of a trial is affected by the set size on the *previous* trial. One possibility for this is a preparation effect. It is possible that, following a difficult trial, subjects prepare for another trial of similar difficulty. It is more likely, however, that this reflects the persistence of the pupillary response. Recall that the phasic response tended to increase with memory load; pupil sizes were overall larger following large set sizes and comparatively smaller after small set sizes. If the pupil does not return to resting baseline as quickly as it has previously been assumed (e.g., Beatty, 1986), then the baseline measurement taken on the *next* trial will be related to the set size on the previous trial. This is, in fact, what we observed. The fact that this effect weakens in the feedback+money condition is revealing: In this condition, there was much more time between the end of recall and the beginning of the next trial's baseline measurement. This extra delay was imposed to allow subjects time to view their earnings for the trial. Specifically, the feedback+money condition was at a minimum 3500 ms longer (following recall) than the feedback alone condition. As illustrated in Table 3, the extra delay led to a weakening of the relationship, though it is still present (r = .55 for high spans, r = .29 for low spans, though the latter was not statistically significant). The pupil thus does *not* quickly return to resting levels.

Limitations

As with most physiological measures, pupil diameter is affected by a multitude of factors. We cannot discount the possibility that our data were in some way influenced by such variables as state and trait anxiety, smoking status, caffeine use, and so forth, all of which were not recorded. However, it is unlikely that our data were seriously contaminated by such factors. Most predictions regarding the effects of these variables would be targeted at low spans. In other words, one might argue that low spans suffer more anxiety or that they self-medicate by ingesting caffeine or nicotine. If this were true, then one would expect larger preexperimental tonic pupil size in low spans as compared to high spans. As we have seen, this was not the case.

We must also mention that our use of extreme-groups methodology is not without problems. As many have argued, artificial dichotomies can lead to overestimation of effect sizes and biased power estimates (Preacher, Rucker, MacCullum, & Nicewander, 2005). When used with care, however, extreme-group designs offer much with respect to economy and have become commonplace in differential psychology (Conway et al., 2005).

That aside, the results from the current work make clear that an effort explanation for individual differences in working memory capacity is simply untenable. Stated differently, the commonly observed dissociation between high and low span subjects is not due, in total, to differing levels of effort, though effort surely does play some role. Some other construct must lie at the heart of working memory; exactly what that is awaits conclusive evidence, but the region of possibilities is now a little smaller.

REFERENCES

Ahern, S. K., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*, *205*, 1289–1292.

Beatty, J. (1986). The pupillary system. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: System, processes, and applications* (pp. 43–50). New York: Guilford.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson(Eds.), *Handbook of psychophysiology* (2nd ed, pp. 142–162). New York: Cambridge University Press.

Bourne, P. R., Smith, S. A., & Smith, S. E. (1979). Dynamics of the light reflex and the influence of age on the human pupil measured by television pupillometry. *Journal of Physiology*, *293*,1p.

Christopher, G., & MacDonald, J. (2005). The impact of clinical depression on working memory. *Cognitive Neuropsychiatry*, *10*, 379–399.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, *4*, 577–590.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

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

Engle, R. W., Cantor, J., & Carullo, J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 972–992.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, *77*, 340–355.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resources limitations. *Psychophysiology*, *33*, 457–461.

Granholm, E., & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*, 1–6.

Heitz, R. P., & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General*, *136*, 217–240.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585.

Kahneman, D., & Peavler, S. (1969). Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology*, *79*, 312–318.

Kane, M. J., & Engle, R. W. (2003). Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavioral Research Methods, Instruments, & Computers*, *31*, 429–432.

Larson, G. E., Saccuzzo, D. P., & Brown, J. (1994). Motivation: Cause or confound in information processing/intelligence correlations? *Acta Psychologica*, *85*, 25–37.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*, 368–387.

Peavler, S. W. (1974). Pupil size, information overload, and performance differences. *Psychophysiology*, *11*, 559–566.

Preacher, K. J., Rucker, D. D., MacCullum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192.

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

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

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

Verney, S. P., Granholm, E., & Dionisio, D. P. (2001). Pupillary responses and processing resources on the visual backward masking task. *Psychophysiology*, *38*, 76–83.

(Received October 18, 2006; Accepted July 23, 2007)