Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 3, 615–622

0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.3.615

BRIEF REPORT

Working Memory, Attention Control, and the *N*-Back Task: A Question of Construct Validity

Michael J. Kane
University of North Carolina at Greensboro

Andrew R. A. Conway
Princeton University

Timothy K. Miura and Gregory J. H. Colflesh
University of Illinois at Chicago

The *n*-back task requires participants to decide whether each stimulus in a sequence matches the one that appeared *n* items ago. Although *n*-back has become a standard "executive" working memory (WM) measure in cognitive neuroscience, it has been subjected to few behavioral tests of construct validity. A combined experimental–correlational study tested the attention-control demands of verbal 2- and 3-back tasks by presenting $n - 1$ "lure" foils. Lures elicited more false alarms than control foils in both 2- and 3-back tasks, and lures caused more misses to targets that immediately followed them compared with control targets, but only in 3-back tasks. *N*-back thus challenges control over familiarity-based responding. Participants also completed a verbal WM span task (operation span task) and a marker test of general fluid intelligence (Gf; Ravens Advanced Progressive Matrices Test; J. C. Raven, J. E. Raven, & J. H. Court, 1998). *N*-back and WM span correlated weakly, suggesting they do not reflect primarily a single construct; moreover, both accounted for independent variance in Gf. *N*-back has face validity as a WM task, but it does not demonstrate convergent validity with at least 1 established WM measure.

Keywords: working memory, memory span, *n*-back, intelligence, individual differences

Almost a century after Jacobs (1887) invented memory span as a test of intellect, psychologists made an important psychometric advance with the development of *complex* or *working memory* (WM) span tasks (Daneman & Carpenter, 1980). Relative to *simple* span tasks, which ask participants to immediately recall short stimulus lists, complex span tasks better reflect WM as a system involving both storage and control processes that maintain access to information in the service of complex cognitive activities (Baddeley, 1986). They do so by requiring participants to remember stimulus sequences during an ongoing secondary task. Reading span, for example, presents to-be-recalled items, such as words, in alternation with sentences to comprehend; operation span presents memoranda in alternation with equations to verify (Conway et al., 2005). Remembering in these tasks is thus challenged, as it often is in life, by periodically switching focal attention away from the to-be-remembered information (Barrouillet, Bernadin, & Camos, 2004). Complex span tasks have become central to WM theory, not only because they reflect ecologically valid memory demands but also because their performance predicts variation in general cognitive abilities. Indeed, WM span may account for half the normal variability in general fluid intelligence (Gf; Conway, Kane, & Engle, 2003; Kane, Hambrick, & Conway, 2005).

As neuroimaging technologies, such as functional magnetic resonance imaging (fMRI), developed rapidly in the 1990s, the *n*-back task, rather than complex span, became the dominant measure in investigations of the neurological substrates of immediate memory (Owen, McMillan, Laird, & Bullmore, 2005). *N*-back tasks are continuous-recognition measures that present stimulus sequences, such as letters or pictures; for each item in the sequence, people judge whether it matches the one presented *n* items ago. *N*-back has face validity as a WM task because participants must maintain and update a dynamic rehearsal set while responding to each item.

The ascendancy of *n*-back is nonetheless striking in contrast to WM span, because *n*-back has received little empirical validation as a WM measure. WM span has been extensively validated against other immediate-memory measures, for example, by showing stronger relations to memory tasks requiring information manipulation than to those demanding mainly rehearsal (e.g., Engle, Tuholski, Laughlin, & Conway, 1999). Furthermore, WM span

Michael J. Kane, Department of Psychology, University of North Carolina at Greensboro; Andrew R. A. Conway, Department of Psychology, Princeton University; Timothy K. Miura and Gregory J. H. Colflesh, Department of Psychology, University of Illinois at Chicago.

Timothy K. Miura is now at the Department of Psychological and Brain Sciences, Indiana University.

Portions of this work were supported by Air Force Office of Scientific Research Grant F49620-00-1-131. We are grateful to Maren Iverson and Candice Panergalin for their assistance in data collection.

Correspondence concerning this article should be addressed to Michael J. Kane, Department of Psychology, University of North Carolina at Greensboro, P.O. Box 26170, Greensboro, NC 27402-6170, or to Andrew R. A. Conway, Department of Psychology, Green Hall, Princeton University, Princeton, NJ 08544-1010. E-mail: mjkane@uncg.edu

predicts abilities thought to involve WM processes but not those that should be WM-free (Engle & Kane, 2004). Finally, WM span has seen myriad task analyses and experimental manipulations to clarify the processes driving its performance and predictive power (e.g., Conway & Engle, 1996; Engle, Cantor, & Carullo, 1992; May, Hasher, & Kane, 1999; Unsworth & Engle, 2006). In stark contrast, *n*-back has rarely been compared with other WM tasks or used to predict cognitive ability, it has stimulated little experimental or parametric work to illuminate its psychological properties, and its neural activation profile has rarely been compared with other WM tasks within participants. Below we review the few behavioral investigations of *n*-back's validity.

Task Analyses and the Validity of *N*-Back

A commonly asked empirical question about *n*-back is whether it relies primarily on domain-specific or general processes (e.g., Nystrom et al., 2000), but studies examining how different stimulus sequences affect performance are more relevant to its validity as a WM task. In particular, if WM capacity reflects the attentional control of interference (e.g., Engle & Kane, 2004; May et al., 1999), then *n*-back sequences that strongly elicit such interference should affect performance. For example, 1-back *lure* matches presented within a 2-back task (e.g., *A–B–**C–C***), and 3-back lure matches (e.g., ***A**–B–C–**A***), should elicit strong propensities to respond in error because of their familiarity and resemblance to targets. In contrast, 1-back lures in a 4-back task, and 4-back lures in a 1-back task, should have weaker familiarity effects. Unfortunately, although some *n*-back studies include $n + x$ or $n - x$ lures in their design, they rarely provide details about lure probabilities or analyses of lure-trial performance (but see Gray, Chabris, & Braver, 2003; McElree, 2001; Oberauer, 2005).

However, a series of *n*-back experiments from the 1960s provided preliminary evidence for the lure effects predicted by some WM theories. Moore and Ross (1963; see also Ross, 1966a, 1966b) tested participants in a 2-back task consisting of $n - 1$ and $n + 1$ lures. The stimulus sequences also presented postlure targets, that is, targets immediately following $n - 1$ lures. Both $n + 1$ and $n - 1$ lures elicited more false alarms than did control foils. Moreover, targets following 1-back lures yielded more misses than did targets following $n + 1$ lures or control foils. Ross's data thus suggested that $n - 1$ matches challenge attention control over familiarity-based interference.

The limited research comparing *n*-back with other putative WM tasks makes a more mixed case for its validity. On the negative side, a composite of 2- through 5-back performance correlates more strongly with simple *short-term memory* span than it does with complex WM span (Roberts & Gibson, 2002); as well, *n*-back sometimes accounts for identical variance in language comprehension as does simple span (Kwong See & Ryan, 1995), and completely different variance than does complex span (Roberts & Gibson, 2002). On the positive side, with respect to intelligence and achievement, 2-back latencies decrease with increasing IQ (Gevins & Smith, 2000; Hockey & Geffen, 2004), and *n*-back false alarm rates correlate negatively with teacher ratings of children's academic performance (Aronen, Vuontela, Steenari, Salmi, & Carlson, 2005). With respect to executive control, *n*-back shares more variance with Stroop performance than does short-term memory span (Kwong See & Ryan, 1995); *n*-back tasks significantly impair eye-movement suppression during simultaneous antisaccade tasks (Mitchell, Macrae, & Gilchrist, 2002); *n*-back scores correlate with teachers' attention ratings (Aronen et al., 2005); and, in outpatients with schizophrenia-spectrum diagnoses, 2-back performance predicts thought disorder (Kerns & Berenbaum, 2003). Although the data are few, *n*-back accounts for some variance in intelligence and attention control. Given the weak relation between *n*-back and complex span, however, *n*-back and WM span may not similarly account for variability in intelligence and attention.

Joint Explorations of *N*-Back Validity and Task Parameters

Two recent studies combined task-analytic examinations of lure effects with tests of *n*-back's construct validity, with mixed results. Oberauer (2005, Experiment 2) tested 120 young adults in *n*-back, four WM capacity tasks (two of which were span tasks), and two memory search/recognition tasks requiring target discrimination against recently presented foils. Twenty-five percent of *n*-back trials presented an *n*-minus lure, thus requiring a similar discrimination. The results provided modest evidence for *n*-back's validity. On one hand, a composite WM capacity measure was uncorrelated with *n*-back latencies and lure false alarms (and each WM task was uncorrelated with *n*-back d' that was based on lure false alarms); the only significant WM effect was in misses. On the other hand, in latent-variable analyses combining the lure discrimination trials across both *n*-back and memory-search tasks, a WM capacity factor predicted 23%–36% of the variance in lure accuracy (d'). Thus, WM capacity shared substantial variance with familiarity-based interference when both were expressed as latent variables derived from multiple tasks (one of which was *n*-back).

A somewhat clearer case for *n*-back's validity was made by Gray et al. (2003), whose 48 participants completed 3-back tasks during fMRI scanning and the Ravens Advanced Progressive Matrices Test (RAPM; Raven, Raven, & Court, 1998), a standardized Gf measure, outside the scanner. Approximately 12% of trials presented 2-, 4-, and 5-back lures. Lure foils yielded lower accuracy and longer response times than did nonlure foils, and RAPM predicted performance on lure and target trials (rs = .36). Moreover, RAPM scores correlated with brain activity elicited by *n*-back in cortical regions involved in attention control (rs $\approx$ .45–.60), and the correlation between RAPM and *n*-back lure accuracy was eliminated after statistically controlling for this brain activity. These data suggest a close association between *n*-back and the brain's WM-control functions.

In summary, Oberauer (2005) suggested weak-to-modest relations between *n*-back and complex WM tasks, whereas Gray et al. (2003) demonstrated a strong relation between *n*-back and RAPM. This literature clearly needs to assess correlations among *n*-back, WM span, and Gf to determine whether the modest association between *n*-back and WM span is accompanied by shared or unique predictions of intelligence. If *n*-back and WM span correlate weakly and account for independent Gf variance, they cannot both be valid measures of a single WM construct. In the present study, then, we examined the effects of lure trials on *n*-back performance and tested the relations among *n*-back, complex WM span, and RAPM.

Method

Participants and General Procedure

One hundred thirty-five undergraduates from the University of Illinois at Chicago participated in partial fulfillment of a course requirement. We omitted 3 participants from analyses because they did not follow instructions, leaving $N = 132$. Participants completed tasks in 3 separate sessions, with *n*-back and the operation span task (OSPAN) completed individually and RAPM in small groups.

N-Back Task

Design. We manipulated three independent variables within participants: memory load (2-, 3-back), sequence type (control, lure), and stimulus type (foil, target), with memory load between blocks and the other variables randomly within blocks. Control foils did not match the letter 5-, 4-, 2-, or 1-back in the sequence for 3-back blocks, or 4-, 3-, or 1-back in 2-back blocks. In contrast, lure foils matched a recent letter in the sequence but not the letter *n*-back (e.g., the second *B* in the sequence *R–B–B* is a 1-back lure in 2-back). Three-back blocks presented primarily 2-back lures (e.g., the second *B* in the sequence *Q–B–R–B*); we included only one 1-back lure per list, which was not analyzed. Two-back blocks presented only 1-back lures. Control targets matched the letter *n*-back in the sequence. Postlure targets also matched the letter *n*-back but occurred immediately after a lure (e.g., the third *B* in the sequence *B–B–B* is a 2-back postlure target; the third *B* in the sequence *B–R–B–B* is a 3-back postlure target).

Materials. Eight phonologically distinct letters served as stimuli (*B*, *F*, *K*, *H*, *M*, *Q*, *R*, *X*). Each memory load presented four lists of 48 letters each. Participants performed eight blocks of 48 trials, four blocks per memory load. Each letter appeared 6 times within a list, once as a target. Eight targets (16.67% of trials) and 40 foils (83.33% of trials) appeared per block.

Procedure. Each participant first performed a 2-back and 3-back practice block of 40 trials each, and then eight critical blocks of 48 trials each (alternating between 2- and 3-back). Trials began with a centered fixation cross on-screen for 500 ms, followed by the stimulus in that location for 500 ms, followed by a blank 2,000-ms interstimulus interval. We instructed participants to respond as quickly and accurately as possible whether each letter matched the *n*-back letter. Participants pressed the *1* and *3* numeric keypad keys for "yes" and "no," respectively. To prevent recognition based on perceptual features only, we structured the trials so that letters randomly appeared in either upper or lower case.

OSPAN

OSPAN required participants to solve mathematical operations and remember unrelated words for immediate recall (see Conway et al., 2005; Kane et al., 2004). Each display presented an equation and word together on-screen (e.g., "Is (6/3) + 2 = 4 ? class"). The participant read the problem aloud, said "yes" or "no" to verify the answer, and read aloud the word for later recall. Immediately after an equation-word pair was completed, the experimenter presented the next one. Trials presented 2–5 equation-word pairs before recall. Participants attempted three trials of each size, in the same pseudo-random order. We used a partial-credit, unit-scoring procedure (Conway et al., 2005) in which each trial was scored separately as the proportion of words recalled in the correct serial position, and these trial scores were averaged.

RAPM

In this paper-and-pencil Gf test, each item consisted of a 3 × 3 matrix containing eight black-and-white figures with one missing figure. Among the eight response figures, participants chose the one that best completed the pattern. The score was the total number of items answered correctly (we used only the 18 odd-numbered items; see Kane et al., 2004).

Results

We report nondirectional null hypothesis significant tests with $\alpha = .05$ and partial eta-squared (η^2) as an effect size estimate.

Univariate Analyses of N-Back

We conducted four 2 × 2 analyses of variance, with memory load and sequence type as the independent variables, and target accuracy, foil accuracy, sensitivity, and bias as dependent variables (see Figures 1 and 2). We calculated sensitivity (d_L) and bias (C_L) using formulas recommended by Snodgrass and Corwin (1988) for the application of signal detection theory to logistic distributions:

$$d_L = \ln\{[H(1 - FA)]/[(1 - H)FA]\} \quad (1)$$

$$C_L = 0.5[\ln\{[(1 - FA)(1 - H)]/[(H)(FA)]\}] \quad (2)$$

where ln = natural log, H = proportion of hits, and FA = proportion of false alarms. Hit rates and false alarm rates equal to either 0 or 1 were adjusted by .01. Negative C_L scores reflect a liberal, "yes" bias, and positive scores reflect a conservative, "no" bias.

Foil and target accuracy. Two-back foils were correctly rejected more often than 3-back foils, $F(1, 131) = 5.21$, $\eta^2 = .04$, and control foils were rejected more often than lures, $F(1, 131) = 141.07$, $\eta^2 = .52$; these variables did not interact, $F(1, 131) = 1.11$, $p > .05$, $\eta^2 = .01$. For target accuracy, main effects of memory load, $F(1, 131) = 41.86$, $\eta^2 = .24$, and sequence type, $F(1, 131) = 4.52$, $\eta^2 = .03$, were moderated by a Load × Sequence interaction, $F(1, 131) = 5.12$, $\eta^2 = .04$. Simple effects analyses indicated that postlure targets were detected less often than control targets in 3-back, $F(1, 131) = 7.71$, $\eta^2 = .06$, but not in 2-back, $F(1, 131) < 1$.

Sensitivity and bias. Sensitivity (d_L) was higher in 2-back than in 3-back, $F(1, 131) = 27.17$, $\eta^2 = .17$, and higher on control trials than on lure trials, $F(1, 131) = 50.24$, $\eta^2 = .18$, with no interaction between them, $F(1, 131) = 3.65$, $p > .05$, $\eta^2 = .02$. Lures thus impaired overall sensitivity. For bias (C_L), in contrast, main effects of memory load, $F(1, 131) = 24.55$, $\eta^2 = .16$, and sequence type, $F(1, 131) = 36.06$, $\eta^2 = .22$, were moderated by their significant interaction, $F(1, 131) = 10.15$, $\eta^2 = .07$. Simple effects analyses indicated a stronger effect of sequence type in 2-back, $F(1, 131) = 52.64$, $\eta^2 = .29$, than in 3-back, $F(1, 131) = 6.60$, $\eta^2 = .05$, although both were statistically significant. Participants were thus

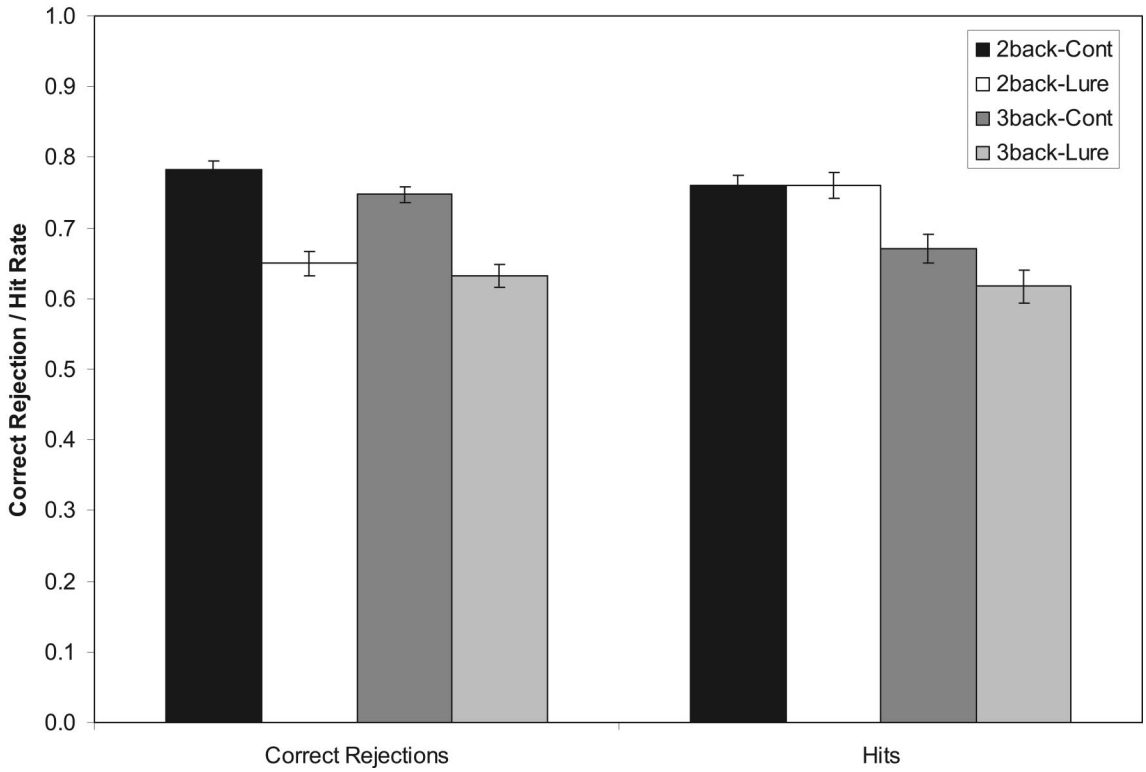

Figure 1. Mean proportion correct responses to foil and target control (Cont) trials and lure trials for the 2-back and 3-back task data ($N = 132$). Error bars represent standard errors.

more conservative in 3-back than in 2-back, and lure sequences elicited a more liberal response bias than control sequences (especially in 2-back).

Multivariate Analyses

Table 1 presents descriptive statistics for all measures used in subsequent analyses. None yielded floor or ceiling effects, and all univariate distributions were normal, as indicated by skew and kurtosis. However, we dropped 3 cases because of large Mahanalobis distance scores (assuming multivariate normality, the probability of obtaining Mahanalobis distances greater than these values is $p < .01$; Mahanalobis, 1936). Therefore, for the following analyses, $N = 129$.

Table 1 also reports reliability estimates for all measures except RAPM and OSPAN. Although we could not estimate reliabilities for RAPM and OSPAN from this sample, previous studies have demonstrated their adequate reliability. Moreover, the correlation between RAPM and OSPAN was equivalent here to a study that used identical tasks (Kane et al., 2004; rs = .33 and .32, respectively) in which both demonstrated good reliability. To calculate reliability estimates for *n*-back, we created two subscales each for accuracy (proportion correct), sensitivity, and bias, one for the mean in Blocks 1 and 3 and the other for Blocks 2 and 4. For each measure, we calculated Cronbach's alpha from the two subscales. As Table 1 demonstrates, these estimates were strong (with the exception of lure C_L), suggesting that *n*-back is a reliable individual-differences indicator of some construct(s).

Table 2 reports correlations among 3-back, OSPAN, and RAPM measures (none of the eight 2-back measures correlated significantly with OSPAN or RAPM). Strikingly, only two of the eight 3-back measures were significantly correlated with OSPAN, and even these were weak (3-back control d_L, $r = .22$; 3-back lure d_L, $r = .17$). These results corroborate prior findings of nonsignificant-to-weak associations between *n*-back and complex span (Oberauer, 2005; Roberts & Gibson, 2002), and they question whether *n*-back measures something similar to complex span. Of importance, our weak correlations were not due to ceiling or floor effects, nonnormal distributions, or lack of reliability. In fact, 3-back performance, although only weakly correlated with OSPAN, was substantially correlated with RAPM (rs = .18–.42; ps < .05).[1]

To further explore the relations among *n*-back, OSPAN, and RAPM—that is, to determine whether our ostensible WM tasks accounted for similar variance in Gf—we conducted a series of regression analyses, first using the raw accuracy measures from *n*-back, and second using signal-detection estimates. In both, we

[1] RAPM correlated negatively with 3-back bias scores (C_L) on targets and foils, indicating that participants with higher RAPM scores had a more liberal response bias than did participants with lower RAPM scores. We are not sure how to explain these correlations, but in any case, our subsequent regression analyses showed that the relation between RAPM and 3-back bias was not significant after accounting for the relation between RAPM and 3-back sensitivity (d_L).

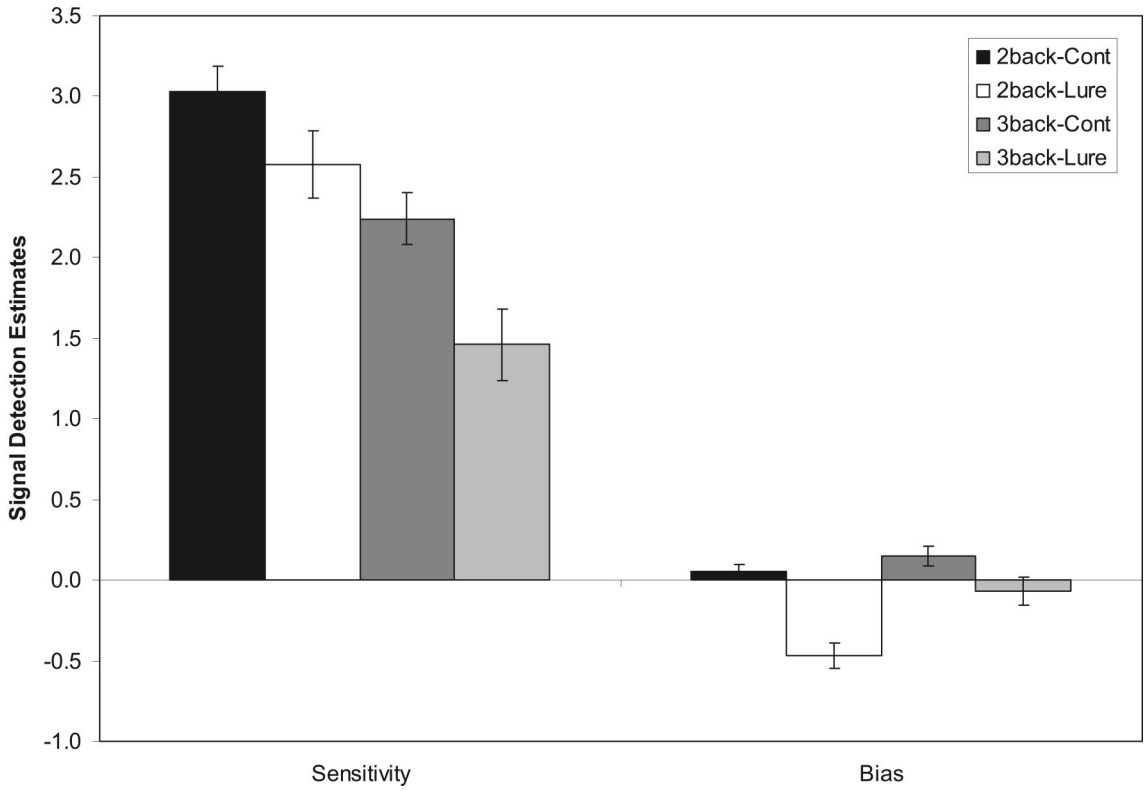

Figure 2. Mean sensitivity (d_L) and bias (C_L) estimates for control (Cont) and lure trials for the 2-back and 3-back task data ($N = 132$). Error bars represent standard errors.

considered only the 3-back data because 2-back did not correlate with either RAPM or OSPAN.

Regressions using 3-back accuracy measures. In their fMRI study, Gray et al. (2003) found not only that 3-back performance correlated with RAPM scores but also that 3-back lure performance accounted for variance in RAPM after controlling for 3-back control performance. We sought to replicate this finding by conducting two hierarchical regression analyses on RAPM, one using 3-back foils and one using 3-back targets (see Table 3). We entered control data in the first step and lure data in the second. (We conducted separate regression analyses for 3-back foils and 3-back targets because multicollinearity among the four predictors prevented us from testing them all in a single model.) Consistent with Gray et al. (2003), 3-back lure performance, on both foil and target trials, accounted for unique variance in RAPM beyond that accounted for by control performance.

Our central question was whether this additional RAPM variance accounted for by 3-back lures would correlate with OSPAN, as one would expect if both reflected similar cognitive-control capabilities. We therefore entered OSPAN into the analysis in a third step. If the shared variance between 3-back lure performance and RAPM is related to OSPAN, then OSPAN should not significantly predict RAPM. As shown in Table 3, OSPAN predicted significant RAPM variance after controlling for 3-back control and lure performance. Each of these measures, then—3-back control, 3-back lure, and OSPAN—accounted for unique RAPM variance.

Regressions using 3-back signal-detection estimates. Control and lure conditions did not differentially correlate with OSPAN, so we averaged them into a composite. Thus, to determine whether *n*-back and OSPAN account for shared or unique variance in RAPM, we conducted multiple regression analyses with OSPAN and 3-back d_L and C_L as predictors (see Table 4). Both OSPAN and d_L accounted for unique variance in RAPM. In fact, subsequent regression analyses (also presented in Table 4), showed that OSPAN's predictive utility was little compromised by including 3-back d_L in the models. OSPAN and d_L thus accounted for mainly

Table 1
Descriptive Statistics for All Measures Used in Multivariate Analyses ($N = 129$)

Variable	*M*	*SD*	Skew	Kurtosis	Reliability
RAPM	10.09	2.96	−.33	.07	
OSPAN	0.64	0.13	.22	−.56	
3-back, control, F	0.75	0.13	−.61	−.23	.80
3-back, lure, F	0.64	0.19	−.53	−.30	.64
3-back, control, T	0.68	0.22	−.75	.01	.84
3-back, lure, T	0.62	0.27	−.54	−.47	.72
3-back, control, d_L	2.24	1.70	.30	.70	.81
3-back, lure, d_L	1.46	2.48	.08	.33	.62
3-back, control, C_L	0.15	0.68	.52	.65	.77
3-back, lure, C_L	−0.07	0.98	.04	.63	.54

Note. RAPM = Ravens Advanced Progressive Matrices Test; OSPAN = operation span task; F = foil accuracy; T = target accuracy; d_L = sensitivity; C_L = bias.

Table 2
Correlations Among All Measures Used in Multivariate Analyses (N = 129)

Measure	1	2	3	4	5	6	7	8	9	10
1. RAPM	—	.33*	.18*	.25*	.39*	.36*	.42*	.36*	−.30*	−.21*
2. OSPAN		—	.15	.14	.14	.13	.22*	.17*	−.08	−.08
3. 3-back, control, F			—	.51*	.25*	.29*	.69*	.43*	.30*	−.05
4. 3-back, lure, F				—	.24*	.29*	.46*	.62*	.02	.20*
5. 3-back, control, T					—	.61*	.82*	.55*	−.79*	−.44*
6. 3-back, lure, T						—	.57*	.88*	−.40*	−.82*
7. 3-back, control, d_L							—	.62*	−.43*	−.33*
8. 3-back, lure, d_L								—	−.29*	−.63*
9. 3-back, control, C_L									—	.37*
10. 3-back, lure, C_L										—

Note. RAPM = Ravens Advanced Progressive Matrices Test; OSPAN = operation span task; F = foil accuracy; T = target accuracy; d_L = sensitivity; C_L = bias.
* $p < .05$.

unique RAPM variance. In contrast, C_L was a significant predictor only when d_L was excluded.

Replicating prior work (Oberauer, 2005; Roberts & Gibson, 2002), then, we found that complex WM span and *n*-back were weakly associated. Surprisingly, these tasks do not appear to be measures of the same construct. We know of no theory that would predict *n*-back and complex span to be unrelated, or that these measures would independently account for individual differences in intelligence, but this is what we found. *N*-back and OSPAN each accounted primarily for independent RAPM variance; alone, OSPAN predicted roughly 11% of RAPM variance, and after *n*-back was accounted for, OSPAN still predicted 7%. Thus, our tests of *n*-back's validity as a WM task provide little evidence that it measures the same executive WM processes engaged by complex span.

Discussion

Combining experimental and correlational methods, we examined executive control in the *n*-back task and asked whether *n*-back trials that should tax control most heavily would correlate most strongly with WM span (OSPAN) and Gf (RAPM) measures. We manipulated *n*-back's control demands by presenting lures that matched the *n*-minus-1-back item. Thus, the task context encouraged restraint over responding to familiar stimuli. We replicated recent findings that such lures increase false alarms relative to control foils (Gray et al., 2003; McElree, 2001; Oberauer, 2005), here in both 2- and 3-back tasks. We also replicated a little-known finding that targets immediately following *n*-minus lures are missed more often than control targets (Moore & Ross, 1963), at least in 3-back.

N-minus-1-back lures challenge attentional control over familiarity-based responding. Do they also engage the memory and executive-control processes elicited by WM span tasks, such as OSPAN? If they do, it is not as much as expected. In a reasonably large, diverse sample, we replicated recent findings of weak correlations between *n*-back and WM span (Oberauer, 2005; Roberts & Gibson, 2002); here the tasks shared only 2%–5% of their variance. Moreover, even though *n*-back and OSPAN both predicted variance in RAPM, they primarily did so independently, with less shared than unique predictive variance between them; even the RAPM variance captured by *n*-back lures was mostly distinct from OSPAN.

Table 3
Summary of Hierarchical Regression Analyses Using OSPAN and 3-Back Foils, or OSPAN and 3-Back Targets, to Predict RAPM Scores (N = 129)

Variable	*B*	*SE*	β	*t*	R^2
Predicting RAPM					
Step 1: Control foils	3.96	1.96	.18	2.02*	.03
Step 2: Lure foils	3.31	1.54	.22	2.15*	.07
Step 3: OSPAN	0.07	0.02	.30	3.58*	.15
Predicting RAPM					
Step 1: Control Ts	5.31	1.13	.39	4.72*	.15
Step 2: Lure Ts	2.18	1.11	.20	1.96†	.18
Step 3: OSPAN	0.06	0.02	.27	3.48*	.25

Note. OSPAN = operation span task; RAPM = Ravens Advanced Progressive Matrices Test; Ts = targets.
† $p < .06$. * $p < .05$.

Table 4
Summary of Regression Analyses Using N-Back Signal-Detection Estimates (N = 129)

Variable	*B*	*SE*	β	*t*	R^2
Predicting RAPM (3 predictors)					
Step 1: OSPAN	0.06	0.02	.26	3.21*	.11
Step 2: 3-back, d_L	0.51	0.15	.32	3.30*	.24
Step 3: 3-back, C_L	−0.34	0.41	−.08	−0.83	.24
Predicting RAPM (2 predictors)					
Step 1: OSPAN	0.06	0.02	.25	3.18*	.11
Step 2: 3-back, d_L	0.58	0.13	.37	4.66*	.24
Predicting RAPM (2 predictors)					
Step 1: OSPAN	0.07	0.02	.31	3.77*	.11
Step 2: 3-back, C_L	−1.13	0.35	−.26	−3.26*	.18
Predicting RAPM (2 predictors)					
Step 1: 3-back, d_L	0.60	0.16	.38	3.85*	.18
Step 2: 3-back, C_L	−0.29	0.43	−.07	−0.68	.18

Note. RAPM = Ravens Advanced Progressive Matrices Test; OSPAN = operation span task; d_L = sensitivity; C_L = bias.
* $p < .05$.

Pragmatically, these findings suggest that theorists cannot necessarily apply the findings from one of these tasks to the other. *N*-back has too long been used by cognitive neuroscientists without serious efforts to assess its construct validity, and now we may have to reappraise past findings. For example, conclusions about the underlying brain circuitry engaged by WM span probably should not be drawn from *n*-back research alone without caution. Fortunately, the arguments for prefrontal cortex involvement in WM span have been based on a convergence of evidence from many different kinds of WM and attention-control tasks (e.g., Kane & Engle, 2002). That said, *n*-back findings have been an important piece of the *prefrontal puzzle* in such work, which may prove to be a problem.

It is less clear what these findings mean for WM theory. Whether theories of WM (and its measurement) focus more on the maintenance of information in the face of simultaneous processing (e.g., Baddeley, 1986; Just & Carpenter, 1992), or on the control over interference and conflict (e.g., Engle & Kane, 2004; May et al., 1999), or on the simultaneous, coordinative binding of multiple stimuli to their contexts or each other (e.g., Oberauer, 2005), they all predict *n*-back and WM span tasks to measure largely the same thing, that is, to reflect primarily the same WM construct. Why don't they?

We might consider more closely the retrieval demands of each task. Complex span typically demands *serial recall*, whereby participants retrieve items using only self-generated cues. In contrast, *n*-back typically demands *recognition*, whereby participants discriminate target items from familiar foils. Our findings suggest that these two aspects of remembering under interference are only minimally related to one another at an individual-differences level, despite their both being important to Gf variation. Indeed, recognition tests tend to minimize interference in paired-associate learning tasks compared with free- or cued-recall (see Anderson & Neely, 1996), and interference often results in failures to recall anything, rather than simply creating discrimination failures regarding multiple retrieved episodes (e.g., Postman, Stark, & Fraser, 1968). Moreover, *n*-back and similar recognition tasks simultaneously tap both familiarity- and recollection-based processes, with familiarity obscuring the relation to recall-based complex span tasks (Oberauer, 2005). Neuroimaging studies also indicate that recognition tasks requiring participants to exert control in rejecting familiar foils activate more ventral prefrontal areas than those (more dorsal areas) that are most closely associated with WM capacity and other aspects of executive control (e.g., Jonides, Smith, Marshuetz, & Koeppe, 1998). It appears as though control over memory-discrimination processes, such as those captured by dynamic memory-updating tasks (e.g., Miyake, Friedman, Emerson, Witzki, & Howerter, 2000), might be only loosely related (behaviorally, functionally, and neuroanatomically) to control over retrieval-under-interference processes.

Of potential importance, a recent individual-differences study found that an *n*-back *recall* task correlated substantially with OSPAN, with *r*s $\approx$ .50 across two samples (Shelton, Metzger, & Elliott, in press). *N*-back was modified from Dobbs and Rule (1989) to present lists of six or eight words and to require participants to recall the word that appeared 0-, 1-, 2-, or 3-back, with no external cues. Although Shelton et al. (in press) included no Gf measure with which to compare *n*-back and OSPAN contributions, their findings suggest that *n*-back captures variance from different constructs depending on the parameters of its embedded memory test. *N*-back reflects similar processes as complex memory span when it demands free recall but different processes from complex span when it demands speeded recognition. We suggest that it is imperative for future work to address this possibility directly.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237–313). New York: Academic Press.

Aronen, E. T., Vuontela, V., Steenari, M.-R., Salmi, J., & Carlson, S. (2005). Working memory, psychiatric symptoms, and academic performance at school. *Neurobiology of Learning and Memory, 83,* 33–42.

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

Barrouillet, P., Bernadin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133,* 83–100.

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.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7,* 547–552.

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

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging, 4,* 500–503.

Engle, R. W., Cantor, J., & Carullo, J. 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., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 145–199). New York: Academic Press.

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.

Gevins, A., & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex, 10,* 829–839.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6,* 316–322.

Hockey, A., & Geffen, G. (2004). The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence, 32,* 591–605.

Jacobs, J. (1887). Experiments on "prehension." *Mind, 12,* 75–79.

Jonides, J., Smith, E. E., Marshuetz, C., & Koeppe, R. A. (1998). Inhibition in verbal-working memory revealed by brain activation. *Proceedings of the National Academy of Sciences, USA, 95,* 8410–8413.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122–149.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intel-

ligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9,* 637–671.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 66–71.

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

Kerns, J. G., & Berenbaum, H. (2003). The relationship between formal thought disorder and executive functioning component processes. *Journal of Abnormal Psychology, 112,* 339–352.

Kwong See, S. T., & Ryan, E. B. (1995). Cognitive mediation of adult age differences in language performance. *Psychology and Aging, 10,* 458–468.

Mahanalobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science, India, 2,* 49–55.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition, 27,* 759–767.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 817–835.

Mitchell, J. P., Macrae, C. N., & Gilchrist, I. D. (2002). Working memory and the suppression of reflexive saccades. *Journal of Cognitive Neuroscience, 14,* 95–103.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100.

Moore, M. E., & Ross, B. M. (1963). Context effects in running memory. *Psychological Reports, 12,* 451–465.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and localizations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *NeuroImage, 11,* 424–446.

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.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). *N*-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25,* 46–59.

Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes in interference. *Journal of Verbal Learning and Verbal Behavior, 7,* 672–694.

Raven, J. C., Raven, J. E., & Court, J. H. (1998). *Progressive matrices.* Oxford, England: Oxford Psychologists Press.

Roberts, R., & Gibson, E. (2002). Individual differences in sentence memory. *Journal of Psycholinguistic Research, 31,* 573–598.

Ross, B. M. (1966a). Serial order effects in two-channel running memory. *Perceptual and Motor Skills, 23,* 1099–1107.

Ross, B. M. (1966b). Serial order as a unique source of error in running memory. *Perceptual and Motor Skills, 23,* 195–209.

Shelton, J. A., Metzger, R. L., & Elliott, E. M. (in press). A group administered lag task as a measure of working memory. *Behavior Research Methods.*

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Unsworth, N., & Engle, R. W. (2006). A temporal-contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language, 54,* 346–362.

Received April 7, 2006
Revision received February 5, 2007
Accepted February 6, 2007 ■