Memory & Cognition
2007, 35 (5), 1136-1150

The validity of "conceptual span" as a measure of working memory capacity

Michael J. Kane and Tina M. Miyake
University of North Carolina, Greensboro, North Carolina

Three experiments tested whether a modified version of the Clustered Conceptual Span task (H. J. Haarmann, E. J. Davelaar, & M. Usher, 2003), which ostensibly requires active maintenance of semantic representations, predicted individual differences in higher-order cognitive abilities better than short-term memory (STM) span tasks or nonsemantic versions of the "Conceptual" task did. Nonsemantic Conceptual tasks presented short word lists clustered by color, first letter, or initial vowel sound, and cued subjects to recall only 1 of 3 clusters from each list; the Semantic task clustered words by taxonomic category. The Semantic Conceptual task generally failed to predict incremental variance in either verbal abilities or general fluid intelligence beyond the other Conceptual tasks or STM span tasks. Although the Semantic task showed a stronger relation to working memory span tasks than did the nonsemantic tasks (Experiment 3), that stronger relation did not translate into strong prediction of cognitive individual differences.

Individual differences in short-term memory (STM) capacity are traditionally assessed with "span" tasks that present short sequences of items, such as digits or words, for the subject to immediately recall in serial order (Jacobs, 1887; Dempster, 1981). Although these tasks are often referred to as "simple" span tasks (e.g., La Pointe & Engle, 1990; Turner & Engle, 1989), they have prompted a host of competing and complex theoretical models of serial-order memory (e.g., Burgess & Hitch, 1999; Farrell & Lewandowsky, 2002; Henson, 1998; Neath & Nairne, 1995; Page & Norris, 1998). Experimental and neuropsychological work also suggests that dissociable storage systems for phonological and lexical-semantic information contribute to a variety of verbal STM tasks (e.g., Crosson et al., 1999; Haarman & Usher, 2001; Hanten & Martin, 2000; Martin & Romani, 1994; Tehan & Lalor, 2000; for a review, see Martin & Freedman, 2001). Based on the idea that a dedicated memory system for maintaining access to lexical-semantic representations might be especially important to successful language comprehension and verbal problem solving, Haarmann, Davelaar, and Usher (2003) developed a new span task, "Conceptual Span," to measure variation in semantic STM and its association to verbal abilities. The present study critically explores the construct validity of Conceptual Span as an individual-differences measure of immediate-semantic-memory capacity.

The Conceptual Span Task

Traditional STM tasks require verbatim repetition of unrelated stimuli in sequential order, and thus encourage subjects to phonologically rehearse the items via inner speech (e.g., Baddeley, Gathercole, & Papagno, 1998; Baddeley, Lewis & Vallar, 1984). In contrast, Conceptual Span was designed to orient subjects' maintenance and retrieval processes toward stimulus meaning. One version of the task presented lists of 9 randomly ordered words that belonged to 3 different semantic categories (e.g., *animals*, *furniture*, and *fruit*), and another "clustered" version presented lists of 12 semantically clustered words, with 4 consecutive words from each category. In both, subjects immediately recalled only the words from one cued category (e.g., *furniture*) in any order, thereby minimizing the importance of serial order and the contribution of phonological processing, and maximizing the role of semantic processing.

Haarmann et al. (2003) found that Conceptual Span predicted variation in reasoning and comprehension more strongly than did STM span tasks which, because they presented either words or pronounceable nonwords, engaged primarily phonological processes. Thus, semantic STM did appear to support complex verbal behavior. Moreover, and perhaps most surprisingly, Conceptual Span predicted verbal individual differences as strongly, if not more strongly, than a Reading Span task did. Reading span is a "complex" working memory (WM) span task that requires subjects to memorize short stimulus sequences that are interpolated with a secondary task, such as reading sentences. WM span tasks correlate robustly with both verbal and nonverbal intellectual abilities (e.g., Conway et al., 2005; Daneman & Merikle, 1996; Kane, Hambrick, & Conway, 2005). It is therefore noteworthy that Haarmann et al. (Study 2) found Conceptual Span to correlate with verbal comprehension and problem-solving

M. J. Kane, mjkane@uncg.edu

tests with *r*s = .34–.51, compared to Reading Span *r*s = .22–.38, and in a pilot study, Conceptual Span correlated with a test of nonverbal reasoning more strongly than WM span tasks usually do, indicating the utility of Conceptual Span to investigating general fluid intelligence (Gf).

Strong predictive validity for Conceptual Span compared to WM span has important theoretical and practical implications. As Haarmann et al. (2003) noted, several views of WM capacity argue that individual differences in WM span, and their considerable correlations with Gf, result primarily from individual differences in domain-general attention processes that are captured by the dual-task requirements of WM span tasks, rather than from some more domain-specific storage or rehearsal processes (e.g., Barrouillet, Bernardin, & Camos, 2004; Cowan, 2005; Engle & Kane, 2004; Lustig, May, & Hasher, 2001). Thus, on one hand, Conceptual Span's apparently broad predictive power may suggest that its cued- and selective-recall requirements, coupled with its de-emphasis of phonological rehearsal, somehow draw upon similar attention-control processes as do WM span tasks. If so, then semantics, per se, may not be important to Conceptual Span as an individual-differences measure. Instead, it could be that any task asking subjects to selectively recall a subset of list items based on some dimension—even a nonsemantic one (such as color, shape, or orthography)—would yield similar patterns of ability correlations. Although such a result would limit Conceptual Span's value to illuminate the nature or functions of a *semantic* WM system, it would not diminish the task's importance. On the contrary, because current WM span tasks all require a secondary processing or manipulation task in addition to memory storage, they are far from process pure and they are cognitively opaque to the researchers who wish to understand them (see Oberauer, 2005). Thus, an ostensibly simpler, "storage only" WM span task would be very useful to the field.

On the other hand, as Haarmann et al. (2003) suggest, Conceptual Span's impressive predictive validity may indicate that the maintenance of semantic representations serves as an additional contributor to higher-order cognitive ability, complementary to the attention-control variance tapped by WM span. If so, this would be an exciting development that could increase, many-fold, the variance in intellectual abilities (including verbal ability and Gf) that is captured by immediate-memory tasks. Latent variables that are statistically derived from the common variance among multiple "complex WM" span tasks account for about 50% of the variance in Gf among young adults (see Kane et al., 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005). Thus, any incremental Gf variance that could be accounted for by multiple "semantic WM" tasks would narrow considerably the search for mechanistic explanations for general fluid ability.

Which interpretation of Conceptual Span is more likely to be correct? Does it reflect attention control in a similar manner to WM span, or does it measure a different construct that is critically linked to maintaining *semantic* memory representations? And, in either case, does Conceptual Span really have the broad predictive power that the initial work by Haarmann et al. (2003) suggests? We designed the present studies to answer these questions, but we can also turn to a few studies using quite similar tasks for preliminary guidance.

Other "Conceptual" Tasks

Several immediate-semantic-memory tasks have good predictive validity.[1] Horn, Donaldson, and Engstrom (1981) created "conceptual" tasks based on experimental paradigms (Bousfield & Bousfield, 1966; Mandler, 1968). In one, subjects categorically sorted 36 words and recalled them immediately or after a delay. In another, subjects heard and immediately recalled 20-word lists comprised of 5 categories each, with the words presented either randomly or clustered. Both tasks correlated with Gf in a sample of 20- to 60-year-olds, and both dramatically reduced the negative effect of age on Gf when they were partialed out: Gf decreased by 3.75 IQ units per decade of age overall, but when the influence of either "conceptual" task was removed, Gf decreased by less than 2 IQ units per decade. These findings suggest that immediate memory tasks requiring selective recall of semantically related items draw on similar cognitive processes as do fluid reasoning tests. But which processes contribute? Horn et al. defined Gf via a composite of letter series, nonverbal matrix completion, and nonverbal topology tests. It is therefore not obvious what contribution *semantic* representations could make. To us, then, these findings suggest that Conceptual Span's correlations with at least some ability tasks derive from domain-general (attention?) processes rather than semantic ones.

Further evidence that Conceptual Span draws at least some of its predictive power from nonsemantic sources comes from research on a cued-recall test of immediate memory developed by Cohen and Sandberg (1977, 1980). They presented lists of 9 digits at a 4/sec rate, visually or aurally, and then immediately cued the subject to recall only the lists' first, second, or third cluster of 3 digits in serial order. Digits have limited semantic representation in memory, and any semantic contribution to performance was further reduced by requiring serial order memory. Even so, in three studies with adolescents, Cohen and Sandberg found that cued-recall scores, particularly for the final, recency cluster of the lists, correlated with IQ in the .40–.60 range. (Ability correlations for the mid-list clusters averaged in the .20–.30 range.)

Moreover, in college students, Cantor, Engle, and Hamilton (1991) found that recall from both middle and recency portions of digit and word lists correlated with verbal Scholastic Aptitude Test (SAT) scores (*r*s ≈ .30); Bunting (2006) found both list types to predict Gf. These data suggest that, on one hand, *any* cued-recall task, even those that minimize semantic processing, show reasonable correlations with verbal and general cognitive abilities. Indeed, Haarman et al. (2003, Study 3) reported that only performance for the middle and final clusters from their clustered Conceptual Span task correlated significantly with verbal abilities, thus mirroring the findings from the Cohen-Sandberg task. On the other hand, the data from adults suggest some hope for Haarmann and colleagues'

view: Cantor et al. (1991) found their cued-recall tasks to correlate a bit less strongly with SAT scores (rs ≈ .30) than did WM span tasks (rs ≈ .40), and the trends in Bunting (2006) were similar for Gf. Given that Haarman et al. found that Conceptual Span correlated as strongly with verbal ability as did WM span, the lower ability correlations found by Cantor et al. and Bunting may suggest that cued-recall tasks for list position do not capture the same individual differences that Conceptual Span does.

The Present Study

We tested the construct validity of Conceptual Span as an individual-differences measure of immediate semantic memory by correlating a clustered version of the task ("Clustered Conceptual Span") to performance on various memory span and cognitive ability tasks, including verbal ability and Gf. We used the Clustered task, exclusively, on the advice of Haarmann et al. (2003). Recall that the Clustered Conceptual task presented consecutive items from each category, rather than presenting items randomly through the list. Haarmann et al. (Study 3) found that *nonclustered* Conceptual Span predicted incremental variance in some verbal abilities beyond that accounted for by the Clustered task. This finding suggests that some aspects of clustering ability (at encoding, retrieval, or both), in addition to maintenance of semantic representations, contribute to the correlations between nonclustered Conceptual Span and complex cognitive abilities. Of course, such clustering abilities need not be specific to semantic information, and they might involve domain-general attentional processes. Thus, the more multiply-determined, nonclustered task only confounds the investigation of the processes contributing to Conceptual Span performance and its correlations with other tasks.

In order to assess the reliability of Haarmann and colleagues' (2003) findings, we compared the predictive power of a modified Clustered Conceptual Span to traditional (phonological) STM span (Experiments 1–3) and WM span (Experiment 3); the modification involved presenting clusters of different sizes (see Method). If the Haarman et al. findings are replicable, Conceptual Span should predict significant variance in verbal ability, Gf, or both, after STM-span-related variance is removed. It might also account for significant incremental variance in intellectual abilities after WM-span-related variance is removed (or, at the very least, Conceptual Span should correlate about as strongly with cognitive abilities as does WM span).

Of most importance, we also compared Clustered Conceptual Span to *nonsemantic* clustered tasks, which cued subjects to recall a subset of items from each list, but the clustering and cuing involved nonsemantic stimulus dimensions. A Color Conceptual task (Experiment 1) presented list items clustered by their color, and subjects immediately recalled only the words that matched the color of the cue. An Orthographic Conceptual task (Experiments 1–3) presented list items clustered by their initial letter, and subjects recalled the words that matched the letter cue. And, finally, a Phonological Conceptual task (Experiments 2 and 3) presented list items clustered by their initial vowel sound, and subjects recalled the words that matched the vowel-sound cue. If maintenance of semantic representations is behind the predictive utility of the Haarman et al. (2003) "Semantic" Conceptual Span task, then it should correlate only modestly with other, nonsemantic Conceptual tasks, and it should predict incremental variance in complex cognitive abilities beyond that accounted for by nonsemantic tasks.

EXPERIMENT 1

Experiment 1 tested the importance of semantic processing to the predictive validity of the conceptual span task. To do this, we created three "conceptual" span tasks that presented words that were clustered according to their semantic category (the *Semantic Conceptual* task), their initial letter (the *Orthographic Conceptual* task), or the color in which they were presented (the *Color Conceptual* task). We also tested the association between conceptual span and simple STM span, as well as their respective powers to predict individual differences in verbal ability and Gf. Three of the 4 STM span tasks presented either abstract words or pronounceable nonwords in order to emphasize phonological rather than semantic or "conceptual" processing.

Method

Subjects

Eighty-seven students at the University of North Carolina at Greensboro (UNCG) participated for credit toward a course requirement; we dropped the data from 7 subjects who did not complete the entire study, leaving data from 80 subjects in all analyses.

Materials

In all the memory tasks, subjects saw the same list of stimuli in the same fixed order.

Clustered conceptual span. We constructed all conceptual tasks from the same pool of 72 words, representing 6 semantic categories (*Animals*, *Produce* (*fruits & vegetables*), *Occupations*, *Tools*, *Clothing*, *Body Parts*; Battig & Montague, 1969) and 6 initial letters (*P*, *S*, *C*, *B*, *L*, *M*). In the Color task, words were presented in red, blue, green, yellow, pink, or white against a black background; in the other conceptual tasks, all words appeared in white against black. Each test comprised 18 trials, each presenting 12 words followed by a cluster recall cue (i.e., a word indicating the to-be-recalled semantic category, first letter, or color, all of which had been presented consecutively). Sound psychometric practice calls for creating tests with items that vary in difficulty (e.g., the application of item-response theory or Rasch models to psychological tests requires variation in item difficulty; e.g., Hambleton, Swaminathan, & Rogers, 1991); in span tasks this is accomplished by presenting lists of varying lengths. In our Conceptual tasks, then, the 12 words of a trial always consisted of a 3-word, a 4-word, and a 5-word cluster in an unpredictable order, and subjects were sometimes asked to recall the 3-word cluster, and other times the 4- or 5-word cluster. Words within a cluster were always unrelated along the dimensions irrelevant to the task (e.g., in the Color task, the words within a color cluster were all represented different semantic categories and all began with different letters).

Within each test, each of the 6 cluster-recall cues appeared three times, but only once to cue each cluster (serial) position. For example, in the Semantic task a cluster of 3 target *Tool* words appeared once in serial position 1, a cluster of 4 target tools appeared once at position 2, and a cluster of 5 target tools appeared once at position 3. Cluster size, cluster category, and cluster serial position were counterbalanced in each task, with each cluster size appearing six times

at each serial position, and each category appearing three times in each serial position (once as a 3-word, 4-word, and 5-word cluster). Each of the 72 words appeared once in each test as a to-be-recalled target word and two times as a nontarget. Three practice trials of 12 words each used the same words as the real trials, but not repeating any of the real-trial target clusters.

Simple STM span. All the STM span tasks presented stimuli in black font against a white background.

Concrete word span. This task consisted of 87 one- and two-syllable words (6 were for practice trials) from Battig and Montague (1969), 12 of which appeared in the Conceptual tasks, plus another 12 drawn from the same semantic categories; the remaining 57 words were drawn from other categories (examples from this task are *goat*, *hill*, *fork*, *brick*, *lawyer*). Eighteen trials presented 3 trials each at list lengths 2–7 (three practice trials presented 2 words each). Although multiple words from the same semantic category could appear within a trial, none of the 12 words from the Conceptual tasks appeared together in a trial. (The sixth trial from this task was deleted from data analyses because it presented a word that had appeared in an earlier trial.)

Concrete nonword span. This task consisted of 87 pronounceable, 1- and 2-syllable nonwords (6 for practice trials), created from real words by changing 1 or 2 letters from words drawn from Battig and Montague (1969), but not appearing in Concrete Word span. All other details matched those of Concrete Word span. Some example nonwords from the test are *yan*, *tace*, *dallen*, *rilo*, and *choan* (derived from *yen*, *tick*, *mallet*, *polo*, and *stone*, respectively).

Abstract word span. This task consisted of 87 (6 for practice trials) 1-syllable words drawn from the MRC Psycholinguistic Database (www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm). All the words had relatively low Concreteness values (200–400, out of a possible 700) as reflected by a combination of published norms (see Coltheart, 1981). Example words are *myth*, *pact*, *rate*, *week*, and *view*. Other details matched those of Concrete Word span.

Abstract nonword span. This task consisted of 87 pronounceable nonwords (6 for practice trials), created from real words by changing 1 or 2 letters from words drawn from the MRC database with the same characteristics as those from the Abstract Word span task. Some examples are *naid*, *balt*, *reez*, *lurst*, and *mot* (derived from *gait*, *tale*, *peep*, *curse*, and *woe*, respectively). All other details matched those of Abstract Word span.

Verbal ability tests. We used three standardized tests adapted by Kane et al. (2004). The Reading Comprehension test was abridged from the Air Force Officer Qualifying Test (AFOQT; Berger, Gupta, Berger, & Skinner, 1990). Subjects had 9 min to complete 14 items. Each item presented a short, incomplete paragraph and subjects chose the best of 5 available completions for it. The modified Analogies test (also from the AFOQT) presented relatively high frequency words to make the test more sensitive to reasoning than vocabulary differences. Subjects had 5 min to complete 18 items that presented an incomplete analogy with 5 possible completions. The Inferences test came from the Educational Testing Service (ETS) Kit of Standard Referenced Tests (Ekstrom, French, Harman, & Derman, 1976). Subjects had 6 min to complete the 10 items from Part 1 of the original test. Each item presented a 1–3 sentence passage and subjects chose the best of 5 inferences that could be drawn without assuming any additional information.

Gf/matrix reasoning tests. We assessed Gf with abridged versions of three matrix reasoning tasks (from Kane et al., 2004), all of which presented patterns of abstract figures with one figure missing. Subjects chose the best available completion. The Ravens Advanced Progressive Matrices (RAPM; Raven, Raven, & Court, 1998) presented 18 items with a 25-min time limit. The matrix reasoning test from the Wechsler Abbreviated Scale of Intelligence (WASI; The Psychological Corporation, 1999) presented 14 items in 8 min, and the matrix reasoning test from the Revised Beta Examination—Third Edition (BETA III; Kellogg & Morton, 1999) presented 20 items in 4 min.

Procedure

We individually tested subjects in two 2-h sessions. In the first, they completed the seven span tasks in this order: Concrete Word, Abstract Nonword, Color Conceptual, Orthographic Conceptual, Semantic Conceptual, Concrete Nonword, and Abstract Word. In the second, subjects completed the three Gf and three verbal ability tasks in this order: RAPM, Reading Comprehension, WASI, Analogies, BETA, and Inferences. All tasks provided written instructions read aloud by the experimenter; all tasks required written responses. Subjects began the next ability/reasoning task only after the full time was used for the previous one.

All conceptual span tasks presented each word in lowercase font for 1 sec, with a 250-msec blank screen between each. A trial was terminated by the appearance of an uppercase word and question mark as a cluster-recall cue (i.e., the category name for the Semantic task (*CLOTHING?*), the first letter for the Orthographic task (*L?*), or the color name (*BLUE?*), presented in that color, for the Color task). Subjects read each word silently and recalled the cued cluster of items in any order; instructions emphasized the unimportance of serial recall and the importance of recalling items from the cued cluster only. We imposed no time limit on recall. Each conceptual task began with three practice trials of 12 items each. Before practice in the Semantic and Orthographic tasks, subjects saw the full list of stimuli, categorized by the relevant dimension for the task (e.g., by taxonomic category for the Semantic task), and read each word aloud. This was done to familiarize subjects with the categories and their exemplars, as well as to help induce enough proactive interference so that subjects would rely on STM, rather than LTM during the test (see Haarmann et al., 2003); PI was also induced through subjects' prior exposure these words in the Color task (and to some in the Concrete Word task).

All STM tasks presented stimuli in lowercase font for 1 sec, with a 500-msec blank screen between each. Each trial was terminated by a recall cue on-screen ("*???*"). Subjects read each stimulus silently and provided written recall in serial order; we encouraged subjects to guess or leave a blank space if they could not remember an item at a given serial position. Recall time was unlimited. Each STM task began with three practice trials of two items each.

Data Scoring

For span tasks, the total score reflected the mean proportion correct, averaged across all the trials in the task. In the STM tasks, we considered a stimulus within a trial to be correct only if it was recalled in its appropriate serial position. For example, in a list-length-3 trial, such as *truth-last-aim*, an output of "truth, aim, last" or "_____, last, game" would yield a trial score of .333, because only one word was recalled in serial position. The nonword tasks required each stimulus to be spelled perfectly; the word tasks required that the word unambiguously represent the presented word. Serial order did not enter into scoring the Conceptual tasks. For all the ability and reasoning tasks, the total score reflected the proportion of correct items.

Results

For all inferential statistics reported hereafter, alpha = .05; effect sizes are reported as partial eta squared (η_p^2).

Descriptive Statistics and Mean Performance

Table 1 presents descriptive statistics for all measures. All were normally distributed as indicated by skew and kurtosis (with values less than 3 and 4, respectively; Kline, 2004), and the Conceptual and STM span tasks demonstrated adequate reliability as indicated by Cronbach's α. Reliability estimates were not available for the verbal-ability or Gf tests, but they showed good reliability in a

Table 1
Descriptive Statistics for Memory Span, Verbal Ability, and Gf Tasks From Experiment 1 ($N = 80$)

Task	*M*	*SD*	Skewness	Kurtosis	α
Semantic conceptual	0.576	0.124	0.118	−0.705	.790
Orthographic conceptual	0.528	0.123	0.189	−0.598	.810
Color conceptual	0.375	0.101	0.408	−0.090	.648
Concrete word	0.764	0.090	−0.125	0.222	.712
Concrete nonword	0.433	0.107	0.463	0.314	.735
Abstract word	0.740	0.089	0.187	0.399	.735
Abstract nonword	0.372	0.098	0.511	−0.322	.722
Reading comprehension	0.458	0.236	0.774	−0.139	
Analogies	0.555	0.189	0.056	−0.908	
Inferences	0.620	0.217	−0.150	−0.644	
RAPM	0.667	0.169	−0.126	−0.549	
WASI	0.696	0.130	−0.553	0.091	
Beta III	0.781	0.115	−0.486	−0.020	

Note—α, Cronbach's alpha; RAPM, Raven's Advanced Progressive Matrices; WASI, matrix reasoning test from Wechsler Abbreviated Scale of Intelligence; Beta III, matrix reasoning test from the Beta III test.

similar sample (Kane et al., 2004), and their substantial intercorrelations indicated adequate reliability here.

The conceptual tasks differed significantly from each other in *M* performance [$F(2,158) = 182.18$, partial $\eta_p^2 = .70$], with the Color task yielding lower recall than the Orthographic task [$t(79) = 13.16$], and the Orthographic task yielding lower recall than the Semantic task [$t(79) = 4.75$]. We analyzed the STM tasks with a 2 (lexicality: word vs. nonword) × 2 (concreteness: concrete vs. abstract) ANOVA. Word span tasks elicited higher recall than nonword tasks [$F(1,79) = 2{,}266.82$, $\eta_p^2 = .966$], and concrete-word tasks elicited higher recall than abstract-word tasks [$F(1,79) = 51.59$, $\eta_p^2 = .389$]. The lexicality × concreteness interaction was also significant [$F(1,79) = 9.86$, $\eta_p^2 = .12$], but all 4 tests differed from each other (lowest $t = 2.77$, $p < .01$).

Confirmatory Factor Analyses of Memory Span Tasks

Confirmatory factor analyses tested whether all the Conceptual tasks measured some common psychological processes, and at least somewhat different processes than those measured by simple STM span tasks. From inspection of the correlations among tasks in Table 2, both Conceptual and STM tasks showed convergent and discriminant validity, correlating more strongly with tasks of the same type than with the other type (e.g., the Conceptual tasks correlated among themselves with *r*s = .58–.74 and with the STM tasks with *r*s = .39–.66).

We modeled the span data with a 2-factor structure, with the three Conceptual tasks comprising a "Conceptual" factor and the four STM tasks comprising an "STM" factor (we allowed residuals of the two Word STM tasks to correlate and the two Nonword tasks to correlate, given the strong effect of lexicality on recall, described above, and given that we had two tasks of each type). Figure 1 presents the model, which fit the data well (indicated by a *nonsignificant* chi-square, a χ^2/df ratio < 2, an RMSEA index below .10, and CFI and NFI indices above .90) [$\chi^2(11) = 8.34$, $p > .05$, $\chi^2/df = 0.76$, RMSEA = .00, CFI = 1.00, NFI = .98]. Conceptual and STM factors shared more than 60% of their variance, but a model forcing all the span tasks to load on a single factor did not fit the data [$\chi^2(12) = 28.79$, $p < .05$, $\chi^2/df = 2.40$, RMSEA = .13, CFI = .95, NFI = .92].

Regression Analyses

To test whether the Semantic task was a particularly strong predictor of either verbal ability (a *z*-score composite of Reading Comprehension, Analogies, and Inferences) or Gf (a *z*-score composite of Ravens, WASI, and Beta III), we first conducted regression analyses in which the Color and Orthographic tasks were entered as predictors before the Semantic task. If semantic processing is important to complex cognition, then it should account for significant variance in verbal or fluid abilities beyond that accounted for by the nonsemantic Conceptual tasks. We then tested whether *any* of the Conceptual tasks predicted intellectual ability after accounting for simple STM span performance by conducting regressions in which a *z*-score composite measure from the STM tasks was entered before the three conceptual tasks (see Table 3).

All three Conceptual tasks correlated significantly with the verbal ability and Gf composites. With respect to verbal ability, however, the Semantic task did not predict unique variance ($\Delta R^2 = .005$, $p > .05$) after accounting for the Color and Orthographic tasks ($R^2 = .243$). More-

Table 2
Correlation Matrix for Experiment 1 ($N = 80$)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Semantic conceptual		.736	.614	.542	.469	.660	.469	.395	.434	.244	.269	.216	.021
2. Orthographic conceptual			.580	.548	.557	.636	.472	.440	.494	.335	.360	.265	.132
3. Color conceptual				.326	.420	.516	.392	.284	.357	.177	.265	.214	.056
4. Concrete word					.619	.642	.563	.316	.471	.126	.186	.148	.123
5. Concrete nonword						.686	.745	.496	.455	.250	.083	.007	.193
6. Abstract word							.626	.467	.501	.258	.186	.163	.194
7. Abstract nonword								.519	.518	.400	.178	.109	.262
8. Reading comprehension									.661	.609	.322	.291	.406
9. Analogies										.565	.345	.376	.448
10. Inferences											.362	.112	.347
11. Raven's matrices												.489	.467
12. WASI matrix reasoning													.389
13. Beta III matrix reasoning													

Note—WASI, Wechsler Abbreviated Scale of Intelligence.

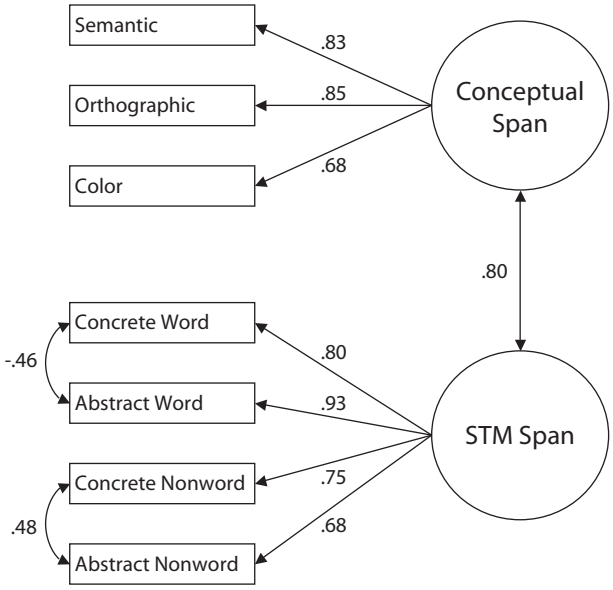

Figure 1. Path model for confirmatory factor analysis of Conceptual and Short-term memory (STM) span tasks from Experiment 1. Paths connecting manifest variables (rectangles) to each other represent correlated error terms added to the model. Paths connecting latent variables (circles) to each other represent the correlations between the constructs, and the numbers appearing on each arrow from a latent variable to a manifest variable represent the loadings of each latent variable onto each task.

over, in the subsequent analysis in which we entered the STM composite first ($R^2 = .291$), *none* of the Conceptual tasks made significant contributions. We found the same pattern in predicting Gf (although, as expected, the verbal memory tasks accounted for less total variance here). The Semantic task accounted for no unique variance after accounting for the Color and Orthographic tasks ($\Delta R^2 = .003$), and none of the Conceptual tasks predicted unique Gf variance after accounting for STM. (Regressions that entered the STM tasks individually, rather than as a composite, yielded identical results except that the Orthographic task predicted incremental variance in verbal ability beyond the STM and Color tasks, $\Delta R^2 = .047$).

Discussion

Does the Semantic Conceptual task measure anything substantially different from nonsemantic versions of this selective-recall task? Experiment 1 suggests not. First, the Semantic, Orthographic, and Color Conceptual tasks formed a tight latent variable in a confirmatory factor analysis, suggesting that they measured largely a single construct. This construct was distinct from, but related to, that tapped by simple span tests of STM, which ostensibly emphasize phonological encoding, rehearsal, and retrieval. Second, the Semantic Conceptual task predicted no variance in either verbal or fluid abilities after accounting for what it shared with the Color and Orthographic tasks. Moreover, none of the Conceptual tasks accounted for incremental variance in intellectual ability beyond that accounted for by simple STM (except where the Orthographic task predicted verbal ability beyond the individually entered STM tasks). So, although Conceptual tasks appear to measure something beyond simple, phonological STM, this unique variance did not seem to predict higher-order cognitive ability.

EXPERIMENT 2

Experiment 1's findings conflicted with Haarmann et al. (2003), and so they required replication. Experiment 2 again contrasted the Semantic Conceptual task with other Conceptual and STM tasks, but here we defined verbal ability and Gf differently, and with more indicators; the verbal tasks were more focused on reading comprehension than on reasoning, whereas the Gf tests assessed both inductive reasoning and spatial visualization. We also replaced the Color Conceptual task with a new task. The Color task differed from the other Conceptual tasks because its target dimension was arbitrary and not a naturally integrated feature of the stimuli. We therefore created a third "integrated" Conceptual task, and one that would be theoretically informative with respect to Haarmann and colleagues' claims about semantic versus phonological STM processes and their respective contributions to verbal ability. Specifically, a *Phonological* Conceptual task presented items clustered by their initial vowel sounds, and so required subjects to selectively recall a subset of list items like the other Conceptual tasks. Because the target cluster dimension was phonological, however, it should correlate more strongly with, and show patterns of predictive validity more similar to, simple STM tasks.

Method

Subjects

One hundred three UNCG students participated for credit toward a course requirement. We excluded data from 16 subjects: 9 did not complete the study, 5 did not understand all the tasks, 1 was extremely sleepy, and several task's worth of data were lost from 1 other subject. Experiment 2 thus included data from 87 subjects.

Table 3
Hierarchical Regression Analyses for Experiment 1 ($N = 80$)

	r	β	F	ΔR^2
Predicting Verbal Composite				
Step 1: Color conceptual	.317	.317	8.70*	.100
Step 2: Orthographic conceptual	.492	.464	14.53*	.143
Step 3: Semantic conceptual	.416	.110	0.50	.005
Step 1: STM composite	.539	.539	32.00*	.291
Step 2: Color conceptual	.317	.074	0.46	.004
Step 3: Orthographic conceptual	.492	.253	3.53	.031
Step 4: Semantic conceptual	.416	−.002	0.00	.000
Predicting Gf Composite				
Step 1: Color conceptual	.225	.225	4.14*	.050
Step 2: Orthographic conceptual	.317	.282	4.53*	.053
Step 3: Semantic conceptual	.212	−.082	0.23	.003
Step 1: STM composite	.224	.224	4.12*	.050
Step 2: Color conceptual	.225	.152	1.46	.018
Step 3: Orthographic conceptual	.317	.270	3.02	.035
Step 4: Semantic conceptual	.212	−.094	0.28	.004

*$p < .05$.

Table 4
Descriptive Statistics for Memory Span, Verbal Ability, and Gf Tasks From Experiment 2 ($N = 87$)

Task	*M*	*SD*	Skewness	Kurtosis	α
Semantic conceptual	0.529	0.131	−0.263	0.231	.833
Orthographic conceptual	0.496	0.114	0.433	0.382	.772
Phonological conceptual	0.243	0.096	0.400	−0.075	.761
Concrete word	0.760	0.094	−0.675	0.809	.780
Abstract word	0.710	0.119	−1.097	1.797	.854
Abstract nonword	0.348	0.105	0.072	−0.366	.795
Reading comprehension	0.314	0.179	1.028	0.731	.760
Nelson-Denny	0.468	0.150	0.298	−0.259	.709
College Board	0.357	0.137	0.363	−0.020	.334
Passage revision	0.421	0.139	0.074	−0.584	.723
RAPM	0.565	0.167	−0.557	0.539	.748
Letter sets	0.638	0.172	−0.277	−0.994	.732
Locations	0.479	0.184	−0.084	0.152	.648
Surface development	0.572	0.252	0.172	−1.219	.942
Paper folding	0.621	0.250	−0.140	−1.145	.800

Note—α, Cronbach's alpha; RAPM, Raven's Advanced Progressive Matrices.

Materials

Memory span tasks. We used five of the span tasks from Experiment 1 (Semantic and Orthographic Conceptual, Concrete Word, Abstract Word, and Abstract Nonword) and created a Phonological Conceptual task to replace the Color task. The Phonological task presented items clustered by their first vowel sound ("EH" as in *vest*, "OW" as in *cow*, "OO" as in *moose*, "EYE" as in *spy*, "OH" as in *bone*, "UH" as in *bucket*; 12 words per cluster type). We clustered by first vowel sounds in order to distinguish the Phonological from the Orthographic task. We also selected items for each vowel sound that differed as much as possible in their spelling of the critical sound (e.g., "EH": *measure*, *leopard*, *pelvis*, *chef*; "OW": *mouth*, *blouse*, *owl*, *gown*; "OO": *moose*, *suit*, *jeweler*, *prune*; "EYE": *nylons*, *coyote*, *pineapple*, *pliers*; "OH": *crowbar*, *bolt*, *coat*, *shoulder*; "UH": *pumps*, *onion*, *banana*, *blood*). Due to vowel-sound constraints, only 25 of the words from the Phonological task matched those from the Semantic and Orthographic tasks (47 were new), but they all represented the same semantic categories (*Animals* (14 words), *Produce* (12), *Jobs* (10), *Tools* (10), *Clothing* (13), *Body parts* (13)). To-be-recalled target items, cluster sizes, and cluster serial positions were counterbalanced across trials as in the other Conceptual tasks.

Verbal reasoning tests. Experiment 2 used 4 verbal tasks emphasizing reading comprehension. AFOQT Reading Comprehension from Experiment 1 was expanded to 20 of the test's original items, with a 10-min time limit. The Nelson-Denny consisted of 6 short passages with 4 questions each, with a 12-min limit. The "College Board" test consisted of 8 passages we selected from PSAT, SAT, and GRE test preparation books. The passages varied in length, and the test consisted of 24 questions with a 15-min limit. Very few subjects completed more than 16 questions, so we scored the test for only these questions (the first 4 questions related to passage 1, the next 6 related to the content of passages 2, 3, or both, and the last 6 related to the content of passages 4, 5, or both). Finally, ACT Passage Revision presented 2 passages and 28 questions, with a 10-min limit. The passages had various words or phrases underlined, and subjects selected among 3 revisions (plus a "no change" option) for the underlined portion that best represented the relevant idea and best fit the style of the passage.

Gf tasks. We measured Gf with five inductive-reasoning and spatial-visualization tasks. The RAPM was identical to that in Experiment 1, but with a 13-min limit. The ETS Letter Sets task consisted of 15 items with a 6-min limit. Each item presented five sets of letters (e.g., *NOPQ*, *DBFL*, *ABCD*, *HIJK*, *UVWX*), of which four sets comprised letters grouped according to the same rule. Subjects selected the "odd-man-out" set that followed a different rule. The ETS Locations test had 14 items with a 6-min limit. Each item presented five rows of dashes grouped together, and in each of four rows, an "X" replaced a dash in one of the groupings according to the same rule. The fifth row presented the numbers 1–5, each replacing one dash, and subjects chose which number would be replaced by an X according to the rule. The ETS Paper Folding test presented 10 items with a 4-min limit. Each item showed a square piece of paper folded 1–3 times and then punched with one or two holes. Subjects selected one of five illustrations representing the positions of the punched holes if the paper were unfolded. Finally, the ETS Surface Development test consisted of 5 items with a 6-min limit. Each item showed a piece of paper that, when folded, would make the 3-dimensional shape next to it. Some edges of the unfolded paper were marked with letters, and five of the edges of the shape were marked with numbers. Subjects matched each of the five numbered edges to the corresponding lettered edge (and so each item required five responses).

Procedure

Subjects completed two 2-h sessions. The first tested subjects individually and the second tested groups of 1–6. In the first session, subjects completed the memory tasks as in Experiment 1, in the following order: Concrete Word, Orthographic Conceptual, Phonological Conceptual, Abstract Nonword, Semantic Conceptual, and Abstract Word. Before beginning the Phonological practice trials, subjects read aloud the stimuli, clustered by initial vowel sound. As in the Semantic and Orthographic tasks, this practice familiarized subjects with the word list, the clustering rules, and the capitalized recall cues (*EH?*, *OW?*, *EW?*, *EYE?*, *OH?*, and *UH?*), and it promoted interference. During the actual Phonological task, the experimenter read aloud each recall cue when it was presented, in order to emphasize its sound.

In the second session, subjects completed the verbal and Gf tests in the following order: College Board Reading Comprehension, RAPM, ETS Surface Development, AFOQT Reading Comprehension, ETS Locations, ETS Letter Sets, Nelson-Denny, ETS Paperfolding, and ACT Passage Revision. Subjects began the next test only after time expired for the previous one.

Results

Descriptive Statistics and Mean Performance

Table 4 presents descriptive statistics for all measures. All were normally distributed as indicated by skew and kurtosis, and the Conceptual and STM span tasks demonstrated adequate reliability. We also calculated reliability estimates for the verbal-ability and Gf tests, and all were adequate except for the College Board comprehension task ($\alpha = .334$); we retained this measure despite its poor reliability estimate, however, because it correlated well with the other verbal measures (rs = .37–.57), suggesting a true reliability of at least .57.

As in Experiment 1, the conceptual tasks differed significantly from each other in *M* performance [$F(2,172) = 464.21$, partial $\eta_p^2 = .84$], here with the Phonological task yielding much lower recall than the Orthographic task [$t(86) = 26.27$], and the Orthographic task again yielding slightly but significantly lower recall than the Semantic task [$t(86) = 3.41$]. The three STM tasks also differed significantly in mean recall [$F(2,172) = 974.60$, $\eta_p^2 = .92$], with the Abstract Nonword task showing lower recall than the Abstract Word task [$t(86) = 35.67$], and the Abstract Word task eliciting a bit worse performance than the Concrete Word task [$t(86) = 4.43$].

Table 5
Correlation Matrix for Experiment 2 ($N = 87$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Semantic conceptual		.745	.591	.413	.496	.434	.312	.382	.067	.355	.239	.278	.150	.330	.229
2. Orthographic conceptual			.645	.469	.451	.453	.360	.366	.109	.360	.100	.235	.079	.237	.174
3. Phonological conceptual				.444	.507	.661	.333	.314	.146	.465	.177	.274	.162	.338	.274
4. Concrete word					.551	.635	.442	.254	.257	.362	.099	.263	.063	.223	.161
5. Abstract word						.650	.245	.232	.234	.225	.018	.148	.070	.024	.074
6. Abstract nonword							.438	.390	.313	.374	.111	.208	.041	.247	.127
7. Reading comprehension								.589	.566	.408	.257	.255	.133	.435	.363
8. Nelson-Denny									.431	.452	.224	.007	.030	.279	.204
9. College Board										.372	.117	−.027	.047	.109	.151
10. Passage revision											.087	.213	.270	.251	.135
11. Raven's matrices												.336	.314	.613	.485
12. Letter sets													.270	.516	.472
13. Locations														.318	.321
14. Surface development															.613
15. Paper folding															

Confirmatory Factor Analyses of Memory Span Tasks

Table 5 presents the correlations among the memory and reasoning tasks. As in Experiment 1, we modeled the Conceptual and STM tasks with two factors. However, because the Phonological Conceptual task should have focused subjects on the items' phonology, performance here should have had more in common with the STM span tasks. Thus, we fit a second model in which both the Conceptual and STM factors loaded onto the Phonological task.

The two-factor model with a link between the Phonological task and the Conceptual factor yielded a .73 correlation between the Conceptual and STM factors (similar to the .80 correlation in Experiment 1), but the model did not fit the data well [$\chi^2(8) = 24.47$, $p < .05$, $\chi^2/df = 3.06$, RMSEA = .16, CFI = .94, NFI = .91]. Allowing the Phonological task to split its loadings between factors, however, yielded an acceptable fitting model, which is presented in Figure 2 [$\chi^2(7) = 12.59$, $p > .05$, $\chi^2/df = 1.80$, RMSEA = .10, CFI = .98, NFI = .96]. This model fit the data significantly better than did the first [$\chi^2_{\text{difference}}(1) = 11.88$]. Note that, here, the Phonological task had equivalent loadings on the two factors, and the correlation between factors was considerably lower than that in Experiment 1, with Conceptual and STM span sharing only 40% of their variance. As in Experiment 1, a model forcing all the span tasks into a single factor did not fit the data [$\chi^2(7) = 53.15$, $p > .05$, $\chi^2/df = 5.91$, RMSEA = .24, CFI = .83, NFI = .81].

Regression Analyses

We again tested for the importance of semantics to the predictive power of Conceptual span by regressing verbal ability and Gf onto the Semantic task after the Phonological and Orthographic tasks (and, also, after the STM tasks). Table 6 presents these analyses. Again, all three Conceptual tasks correlated with verbal ability and Gf (z-score composites of the verbal-ability and Gf scores, respectively). But here, only the Phonological task accounted for significant variance in verbal ability ($R^2 = .164$); the Semantic task predicted no incremental variance ($\Delta R^2 = .005$). As in Experiment 1, none of the Conceptual tasks predicted unique variance in verbal ability in the analysis that first entered the STM composite into the equation (nor did they when they were entered after the individual STM tasks, rather than the composite). Unlike the findings from Experiment 1, however, both the Phonological task (entered first: $\Delta R^2 = .111$) and the Semantic task (entered third: $\Delta R^2 = .04$) accounted for significant unique variance in Gf. Indeed, both of these tasks remained significant predictors of Gf even after the STM composite was accounted for (Phonological $\Delta R^2 = .07$; Semantic $\Delta R^2 = .04$; they were also significant after accounting for the individual STM tasks).

Discussion

Experiment 2 demonstrated that not all "Conceptual" tasks are created equal. The Phonological task, in which

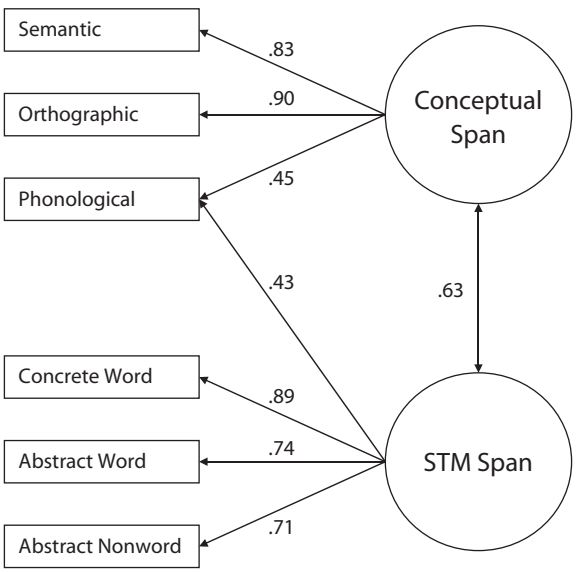

Figure 2. Path model for confirmatory factor analysis of Conceptual and short-term memory (STM) span tasks from Experiment 2.

subjects encoded and retrieved word clusters by phonology, rather than by semantics or orthography, shared as much variance with serial-recall STM tasks as with the other Conceptual tasks. That is, in a confirmatory factor analysis, the Phonological task split its loadings between the "Conceptual" and "STM" factors, suggesting that phonological processing is what distinguishes standard STM tasks from the Semantic (and Orthographic) Conceptual task. Thus, as argued by Haarmann et al. (2003), nonphonological Conceptual tasks do seem to tap some different (nonphonological) processes and abilities than do standard STM tasks. That said, the Phonological task did not behave merely like a standard STM task, for it predicted unique variance in Gf after accounting for the STM-task variance, and variation in phonological STM is often not very predictive of Gf (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999).

But what of the Semantic task? Consistent with the suggestions of Haarmann et al. (2003), but in contrast to Experiment 1, here we found it to predict incremental variance in Gf after STM, Phonological, and Orthographic tasks were accounted for, suggesting that the Semantic task does uniquely tap some mental processes that are important to novel reasoning. Why did the Semantic task "work" here but not in Experiment 1? One might guess that it is because we defined Gf slightly differently here, with less focus on matrix-reasoning tasks and more on general inductive reasoning and spatial visualization. One would be mistaken, however. When the RAPM test, which was common to both experiments, was regressed on the 3 Conceptual tasks from each experiment, the Semantic task accounted for no incremental variance beyond the other 2 Conceptual tasks in Experiment 1 ($\Delta R^2 = .000$), but significant incremental variance beyond the other 2 tasks in Experiment 2 ($\Delta R^2 = .048$). Semantic Conceptual performance thus seems to have a fickle relation with Gf.

We were disappointed that the Semantic task again failed to predict unique variance in verbal abilities after other memory tasks were accounted for. Whereas Experiment 1 had focused more on verbal reasoning, Experiment 2 focused more on reading comprehension, and yet, here, only the Phonological task accounted for unique variance in verbal ability (and only when the model did not include the STM composite; when the model included STM, *none* of the Conceptual tasks accounted for incremental variance). Experiment 2 thus provides only limited support for Haarmann and colleagues' (2003) argument that the Semantic Conceptual task is somehow special in its ability to predict variation in higher-order cognitive ability.

Table 6
Hierarchical Regression Analyses for Experiment 2 ($N = 87$)

	r	β	F	ΔR^2
Predicting Verbal Composite				
Step 1: Phonological conceptual	.405	.405	16.71*	.164
Step 2: Orthographic conceptual	.385	.212	2.71	.026
Step 3: Semantic conceptual	.360	.102	0.45	.005
Step 1: STM composite	.470	.470	24.09*	.221
Step 2: Phonological conceptual	.405	.183	2.28	.020
Step 3: Orthographic conceptual	.385	.140	1.20	.011
Step 4: Semantic conceptual	.360	.048	0.11	.001
Predicting Gf Composite				
Step 1: Phonological conceptual	.333	.333	10.60*	.111
Step 2: Orthographic conceptual	.224	.016	0.01	.000
Step 3: Semantic conceptual	.333	.308	3.94*	.040
Step 1: STM composite	.198	.198	3.46	.039
Step 2: Phonological conceptual	.333	.343	6.80*	.072
Step 3: Orthographic conceptual	.224	.021	0.02	.000
Step 4: Semantic conceptual	.333	.318	4.07*	.042

*$p < .05$.

EXPERIMENT 3

Experiment 3 tested whether the Semantic Conceptual task might better predict versions of verbal-ability tests used originally by Haarmann et al. (2003), the "Pronoun Texts" and "Anomalies" tests (in addition to the Reading Comprehension and Analogies tests we used in the previous experiments). We also sought to replicate either the Experiment 1 or Experiment 2 findings concerning the relation between the Semantic task and Gf, with a computerized version of the RAPM test along with a more verbal Gf test. Experiment 3 also included two WM span tasks in order to test the relations among Conceptual, STM, and WM spans, and to examine their relative contributions to verbal ability and Gf. Finally, based on separate personal communications with Henk Haarmann and Randi Martin (June, 2004), we tested all subjects on the Semantic Conceptual task as the very first task, just in case subjects' prior experience with STM and phonological tasks led them in Experiments 1 and 2 to use some phonological processing, at least some of the time, in the Semantic task.

Method

Subjects

Ninety-five UNCG students participated for partial course credit. We excluded data from 8 subjects: 4 did not complete the study, 1 was a nonnative English speaker, 1 made too many processing errors during Reading span, 1 had participated in Experiment 2, and 1 was repeatedly interrupted by his police radio. Experiment 3 thus included data from 87 subjects.

Materials

Memory span tasks. We used the three Conceptual tasks and the Abstract Word and Abstract Nonword STM tasks from Experiment 2. We also included two "complex," or "WM," span tasks, Operation span (OSPAN) and Reading span (RSPAN). OSPAN required subjects to remember sequences of 2–5 words, in serial order, that were each preceded by an equation to verify (e.g., *Is (10/2) − 3 = 2? DRUM*). Subjects immediately read each equation aloud, then verified whether the answer was correct, and then immediately read the word aloud. The experimenter then immediately presented the next equation-word pair (or recall cue [*???*]) on-screen. Subjects completed four trials of each list length in a pseudorandom order. RSPAN was constructed and run in the same way, but the memoranda were preceded by sentences to judge as "making sense" or not (e.g., *The prosecutor's dish was lost because it was not based on fact. ? LOAFERS*). The memoranda for both OSPAN and RSPAN were drawn from the Conceptual, Concrete Word, and Abstract Word pools, with approximately equal representation. And, for both OSPAN and RSPAN, as in the STM tasks, subjects recalled the words by writing them in serial order; we also scored OSPAN and RSPAN in the same way we scored

Table 7
Descriptive Statistics for Memory Span, Verbal Ability, and Gf Tasks From Experiment 3 ($N = 87$)

Task	*M*	*SD*	Skewness	Kurtosis	α
Semantic conceptual	0.578	0.103	−0.304	0.509	.634
Orthographic conceptual	0.518	0.114	0.419	1.788	.797
Phonological conceptual	0.268	0.096	0.463	−0.045	.749
Abstract word	0.692	0.088	−0.169	−0.296	.766
Abstract nonword	0.347	0.093	0.165	−0.722	.732
Operation span	0.594	0.135	−0.192	0.581	.827
Reading span	0.492	0.153	0.042	0.776	.877
Reading comprehension	0.371	0.209	0.530	−0.188	.797
Analogies	0.396	0.167	0.259	−0.244	.623
Pronoun texts	0.644	0.158	−0.013	−0.734	.563
Anomalies	0.806	0.092	−1.585	3.445	.645
RAPM	0.658	0.089	−0.083	0.464	.578
Following directions	0.422	0.143	−0.273	−0.296	.610

Note—α, Cronbach's alpha; RAPM, Raven's Advanced Progressive Matrices.

the STM tasks (except that retained subjects had to attain ≥85% accuracy in the equation and sentence judgment portions of the WM tasks).

Verbal ability tests. Subjects completed four computerized tests of verbal ability. The AFOQT Reading Comprehension test presented the same 20 items as in Experiment 2, but on a computer screen and requiring keypress responses. Each paragraph appeared without its final sentence for 25 sec (or until the subject finished reading and pressed a key), then replaced by the final sentence and the 5 possible completions with a 20-sec limit. The AFOQT Analogies test presented the same 18 items as in Experiment 1, but with a 12-sec limit for each response. Pronoun Texts (modified from Haarmann et al., 2003[2]) presented 4 passages of 12 sentences each, at a 6 sec per sentence rate (if subjects finished reading before time expired, they hit a key to view the next sentence), with each sentence replacing its predecessor on-screen. After the last sentence disappeared, subjects answered 4 multiple-choice questions about details from the passage (e.g., to whom or what a particular pronoun from the end of the passage referred), each with a 10 sec limit. Finally, the Anomalies test (also modified from Haarmann et al.) rapidly presented 104 sentences, each consisting of 13 words presented for 450 msec each. Following each sentence (at a "RESPOND NOW" screen), subjects had only 1,500 msec to report via keypress whether it made sense or not. Half the sentences were statements and half were questions; half of each made sense (e.g., *What restaurant did the barber that cut the beard recommend to the couple?*) and half were nonsensical (e.g., *Did the housewife that ate the show find the box in the apartment?*). Each trial began with a "READY" screen that prompted the subject to press a key to view the next sentence.

Gf tests. Experiment 3 used a nonverbal and a verbal reasoning test (the verbal test also tapped reading comprehension). A computerized version of RAPM presented 36 items, divided into 3 sets of 12 (with items increasing in difficulty within each set); subjects had 5 min to complete the test. The ETS Following Directions test was divided into 2 parts, with each part presenting 10 questions about a particular 5 × 5 matrix of the digits 1–5 (labeled as Rows A–E and Columns A–E). Each question asked the subject to perform some mental permutations and/or scanning of the digits according to some rule (e.g., *If one number occurs more frequently than another, the answer is the most frequently occurring number; if no number occurs most frequently, the answer is the number appearing in the upper left to lower right diagonal*), and required a keypress response for the correct number. Each problem appeared for 30 sec, with the 5 × 5 matrix of digits in view.

Procedure

Subjects individually completed two 2-h sessions. The first presented the memory tasks in the following order: Semantic Conceptual, OSPAN, Abstract Word, Phonological Conceptual, RSPAN, Abstract Nonword, and Orthographic Conceptual. The second presented the verbal and Gf tests in the following order: ETS Following Directions, Pronoun Texts, RAPM, Anomalies, AFOQT Reading Comprehension, and AFOQT Analogies.

Results

Descriptive Statistics and Mean Performance

Table 7 presents descriptive statistics. All measures were normally distributed as indicated by skew and kurtosis, and the memory tasks demonstrated adequate reliability. Most of the verbal and Gf tasks yielded lower reliabilities than in our previous experiments, but most reliabilities were adequate, and all tests showed reasonable patterns of intercorrelations.

The conceptual tasks again differed significantly in mean performance [$F(2,172) = 373.44$, $\eta_p^2 = .81$], with the Phonological task yielding lower recall than the Orthographic task [$t(86) = 21.28$], and the Orthographic task again yielding lower recall than the Semantic task [$t(86) = 5.03$]. Subjects recalled fewer items in the Abstract Nonword task than the Abstract Word task [$t(86) = 39.71$], and fewer items in the Reading span task than the Operation Span task [$t(86) = 8.78$].

Confirmatory Factor Analyses of Memory Span Tasks

Table 8 presents the correlations among the memory and reasoning tasks. We began with a simple model in which the three Conceptual tasks, the two STM tasks, and the two WM tasks comprised separate factors; this model provided an adequate fit [$\chi^2(11) = 16.85, p > .05, \chi^2/df = 1.53$, RMSEA = .08, CFI = .97, NFI = .92]. However, as in Experiment 2, allowing the Phonological task to split its loadings between Conceptual and STM factors yielded a significantly better fitting model [$\chi^2_{\text{difference}}(1) = 4.56$] presented in Figure 3A [$\chi^2(10) = 12.30, p > .05, \chi^2/df = 1.23$, RMSEA = .05, CFI = .99, NFI = .95]. Replicating Experiment 2, the Phonological task was almost equally associated with the two factors, and the two factors shared almost 40% of their variance. Note also that the Conceptual factor shared about 50% of its variance with the WM factor, while WM and STM shared only about 14% of their variance. Thus, the construct captured by Conceptual span tasks demonstrated considerably strong relations to those tapped by STM and WM span tasks. However, a model with all span tasks loading onto a single factor did not fit the data [$\chi^2(14) = 67.31, p < .05, \chi^2/df = 4.81$, RMSEA = .21, CFI = .74, NFI = .70].

If Haarmann et al. (2003) were right that the Semantic Conceptual task has some properties that are similar to WM span, such as an ability to broadly predict complex cognition (see also Haarmann, Ashling, Davelaar, & Usher, 2005), then one might expect the Semantic task to correlate particularly strongly with WM span tasks. In

Table 8
Correlation Matrix for Experiment 3 ($N = 87$)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Semantic conceptual		.480	.270	.202	.234	.440	.543	.087	.143	.173	.056	.081	.179
2. Orthographic conceptual			.511	.448	.357	.334	.473	.297	.268	.293	.208	.234	.266
3. Phonological conceptual				.413	.454	.248	.345	.302	.379	.364	.352	.156	.299
4. Abstract word					.599	.199	.279	.311	.313	.146	.101	.252	.395
5. Abstract nonword						.223	.298	.323	.387	.278	.237	.213	.369
6. Operation span							.718	.239	.361	.297	.103	.305	.358
7. Reading span								.280	.285	.349	.208	.340	.421
8. Reading comprehension									.647	.232	.382	.181	.247
9. Analogies										.327	.327	.340	.318
10. Pronoun texts											.301	.106	.245
11. Anomalies												.043	.292
12. Raven's matrices													.338
13. Following directions													

fact, inspection of the zero-order correlations reveals that the Semantic task correlated at least as strongly with the two WM tasks (rs = .44 and .53 with Operation and Reading span, respectively) as it did with the two other Conceptual tasks (rs = .27 and .48 with Phonological and Orthographic tasks, respectively). Thus, we fit a final model, illustrated in Figure 3B, in which the Semantic task was allowed to load onto the WM factor. This model fit the data very well, and significantly better than the model depicted in Figure 3A [$\chi^2(9) = 3.92$, $p > .05$, $\chi^2/df = 0.44$, RMSEA = .00, CFI = 1.00, NFI = .98; $\chi^2_{\text{difference}}(1) = 8.38$]. As reflected in the zero-order correlations, then, the model shows the Semantic task loading more strongly onto the WM factor than onto the Conceptual factor.

Regression Analyses

Here we tested whether the Conceptual tasks predicted incremental variance in verbal ability and Gf after accounting for shared variance with STM, WM, or both. In contrast to Experiments 1 and 2, we defined verbal ability and Gf variables as factor scores, rather than as z-score composites. We did so because Experiment 3 used only one commonly accepted Gf marker (RAPM), and although the Following Directions test has face validity as a Gf task, it also involves reading comprehension and, thus, verbal ability. We therefore conducted a principal components analysis, with oblimin rotation, that forced two factors on the reasoning tasks. This yielded a clear verbal reasoning factor, with .62 or greater loadings for AFOQT Comprehension and Analogies, Pronoun Texts, and Anomalies, a .26 loading for Following Directions and a −.16 loading for RAPM. A clear Gf factor also emerged with loadings of .94 for RAPM, .58 for Following Directions, and .37 for Analogies (the latter had also showed a reasonable Gf loading in Kane et al., 2004); all other loadings were <.15. We therefore used the factor scores that were output from these analyses as indicators of each subject's verbal ability and Gf in all regressions. Note that Gf here had more of a verbal "flavor" than it did in Experiments 1 and 2 (with significant variance attributable to Following Directions and Analogies), and so it might be especially likely to show a strong association with the Semantic Conceptual task.

Table 9 presents the results of the regression analyses. Surprisingly, only the Phonological and Orthographic tasks were significantly correlated with the verbal and Gf factors, and so the Semantic task was unlikely to account for any significant incremental variance in the regressions. Indeed, for verbal ability, only the Phonological task accounted for significant variance after STM, WM, or both, were first entered into the models; neither the Orthographic nor the Semantic tasks made significant contributions beyond the variance they shared with the Phonological task (this was true whether the STM and WM tasks were entered individually or as STM and WM composites). With respect to Gf, none of the Conceptual tasks accounted for incremental variance beyond that attributable to STM, WM, or both.

As a secondary consideration, both STM and WM accounted for significant variance in verbal ability and Gf. Regarding the former, STM and WM correlated with the verbal factor with rs = .40 and .37, respectively. WM accounted for a significant 7% of the variance in verbal ability beyond the 16% accounted for by STM. And, in a supplemental analysis in which we regressed verbal ability on WM before STM, STM accounted for a significant 9% of the variance beyond the 13% accounted for by WM. Thus, as in prior work (e.g., Cantor et al., 1991; Engle, Nations, & Cantor, 1990; Engle et al., 1999), WM and STM span tasks accounted for both shared and unique variance in verbal ability. It was a bit unusual, however, that the STM correlations were slightly stronger than the WM correlations (Daneman & Merikle, 1996).

Still more surprising was that STM accounted for considerable unique variance in Gf, given prior findings that verbal STM much less potently predicts nonverbal, general cognitive abilities than does verbal WM (e.g., Conway et al., 2002; Engle et al., 1999; Kane et al., 2004). Here, STM and WM correlated with Gf with rs = .41 and .45, respectively; as shown in Table 9, WM accounted for a significant 12% of the variance beyond the 17% predicted by STM, and in a subsequent analysis, STM accounted for a significant 8% of the variance beyond the 20% predicted by WM. We believe the predictive power of STM here resulted from our defining Gf via both verbal and nonverbal tasks. In fact, if we considered only the RAPM test as our sole Gf marker, WM predicted significant incremental variance in Gf after accounting for STM ($\beta = .30$, $F = 7.89$, $\Delta R^2 = .08$), but STM did not predict sig-

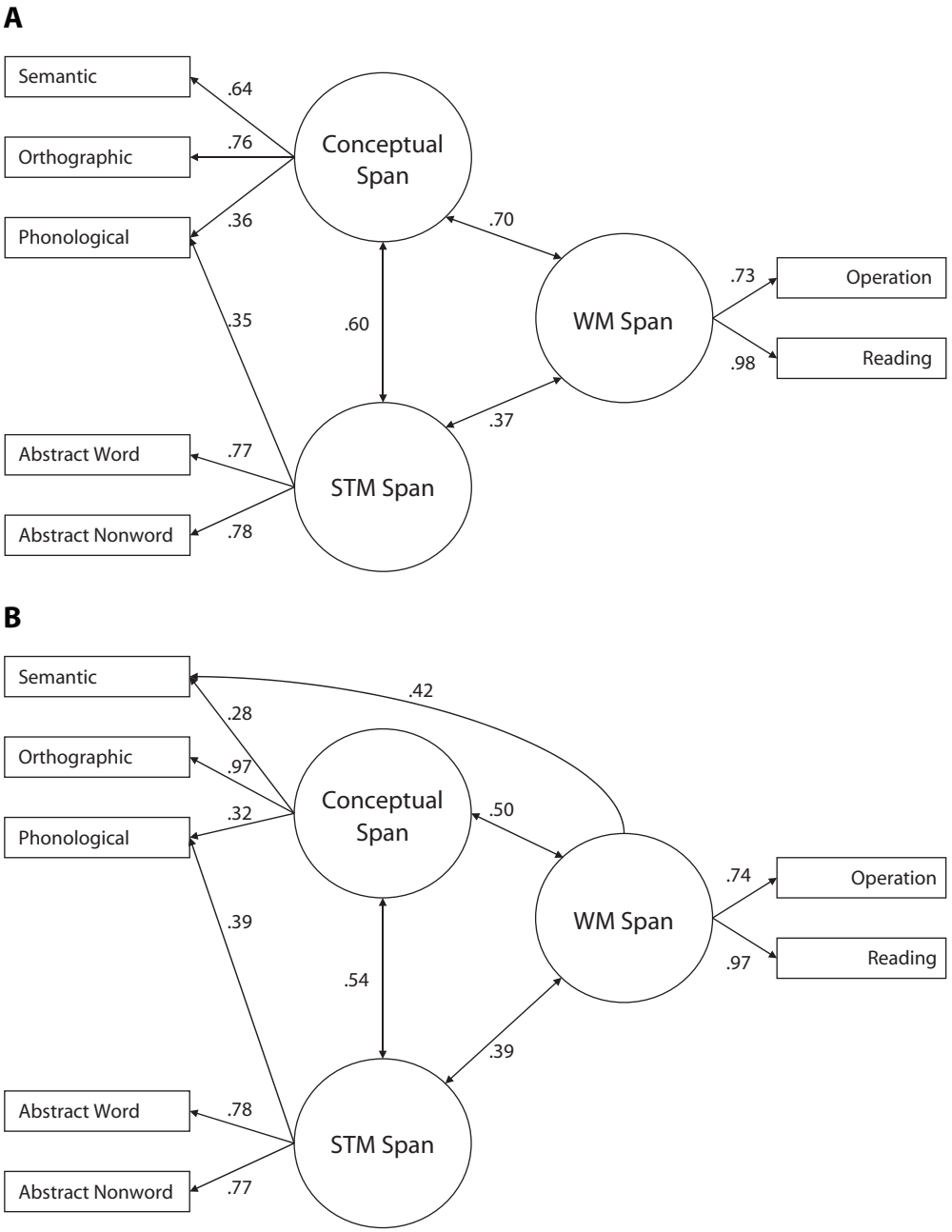

Figure 3. Path model for confirmatory factor analysis of Conceptual, short-term memory (STM), and working memory (WM) span tasks from Experiment 3. Panel A: Model allowing the Phonological Conceptual task to split loadings between the Conceptual and STM span factors. Panel B: A better-fitting model, also allowing the Semantic Conceptual task to split loadings between the Conceptual and WM span factors.

nificant variance in Gf after accounting for WM ($\beta = .17$, $F = 2.51$, $\Delta R^2 = .03$). Thus, as in previous research (e.g., Engle et al., 1999), the shared variance between WM and STM had some power to predict variation in Gf, but WM accounted for significant Gf variance beyond that predicted by STM.

Discussion

Experiment 3 failed to support the claims that Semantic Conceptual span is a better predictor of higher-order cognitive abilities than is simple STM span and an equally good predictor of cognitive abilities as is complex WM span. Even though we used two versions of the Haarmann et al. (2003) verbal criterion tasks here, and even though subjects completed the Semantic Conceptual task before any of the phonological or STM tasks, the Semantic task showed very low correlations with our criterion measures and it accounted for no variation in verbal ability or Gf after the other Conceptual tasks were accounted for. The Semantic task did show a stronger relation to WM span

tasks than did the Orthographic and Phonological tasks, suggesting it tapped something "special" relative to other STM tasks, but this shared variance with WM obviously did not translate into greater predictive validity for higher-order cognition.

Experiment 3 also showed that verbal STM tasks can be strong predictors of some verbal tasks. Although a meta-analysis (Daneman & Merikle, 1996) suggests that prototypical verbal ability tests are usually more strongly associated with WM span than STM span, this is not always the case (perhaps because word knowledge may contribute to both STM span and verbal-ability-test performance; e.g., Engle et al., 1990). Indeed, Haarmann et al. (2003) found word span to correlate as strongly as reading span with several comprehension and verbal problem solving tasks. In contrast, verbal WM seems much more consistently superior to verbal STM in its relation to more domain-general fluid abilities (Conway et al., 2002; Engle et al., 1999; Kane et al., 2004). The apparent equality of STM and WM in predicting Gf here was a function of defining Gf via a combination of verbal reasoning (and comprehension) and nonverbal reasoning tasks. When considering only the nonverbal RAPM as an indicator of Gf, WM accounted for more variance than did STM, and only WM accounted for unique Gf variance.

Table 9
Hierarchical Regression Analyses for Experiment 3 ($N = 87$)

	r	β	F	ΔR^2
Predicting Verbal Composite				
Step 1: Phonological conceptual	.478	.478	25.13*	.228
Step 2: Orthographic conceptual	.349	.142	1.65	.015
Step 3: Semantic conceptual	.154	−.031	0.08	.001
Step 1: STM composite	.396	.396	15.81*	.157
Step 2: Phonological conceptual	.478	.215	12.14*	.106
Step 3: Orthographic conceptual	.349	.091	0.65	.006
Step 4: Semantic conceptual	.154	−.037	0.12	.001
Step 1: WM composite	.366	.366	13.14*	.134
Step 2: Phonological conceptual	.478	.402	16.85*	.145
Step 3: Orthographic conceptual	.349	.061	0.28	.002
Step 4: Semantic conceptual	.154	−.146	1.58	.014
Step 1: STM composite	.396	.396	15.81*	.157
Step 2: WM composite	.366	.271	7.20*	.066
Step 3: Phonological conceptual	.478	.326	9.22*	.078
Step 4: Orthographic conceptual	.349	.022	0.04	.000
Step 5: Semantic conceptual	.154	−.144	1.57	.014
Predicting Gf Composite				
Step 1: Phonological conceptual	.244	.244	5.39*	.060
Step 2: Orthographic conceptual	.289	.223	3.41	.036
Step 3: Semantic conceptual	.144	.003	0.00	.000
Step 1: STM composite	.408	.408	16.94*	.166
Step 2: Phonological conceptual	.244	.060	0.28	.003
Step 3: Orthographic conceptual	.289	.131	1.20	.012
Step 4: Semantic conceptual	.144	−.001	0.00	.000
Step 1: WM composite	.450	.450	21.56*	.202
Step 2: Phonological conceptual	.244	.111	1.19	.011
Step 3: Orthographic conceptual	.289	.079	0.43	.005
Step 4: Semantic conceptual	.144	−.190	2.51	.023
Step 1: STM composite	.408	.408	16.94*	.166
Step 2: WM composite	.450	.359	13.74*	.117
Step 3: Phonological conceptual	.244	−.022	0.04	.001
Step 4: Orthographic conceptual	.289	.001	0.00	.000
Step 5: Semantic conceptual	.144	−.187	2.61	.022

*$p < .05$.

GENERAL DISCUSSION

In 3 experiments we tested whether a new version of the Haarmann et al. (2003) Clustered Conceptual Span task, which ostensibly requires subjects to actively maintain semantic representations, was a better individual-differences predictor of verbal and nonverbal cognitive abilities than were traditional STM tasks or nonsemantic versions of the Clustered "Conceptual" task. We created nonsemantic Conceptual tasks by presenting short word lists clustered by their color (Experiment 1), their first letter (Experiments 1–3), or their initial vowel sound (Experiments 2 and 3), and requiring subjects to selectively recall only one cluster from each list. Across experiments, the Semantic Conceptual task generally failed to predict incremental variance in either verbal ability or Gf after the other Conceptual tasks, or traditional STM tasks, were accounted for. Although the Semantic task accounted for unique Gf variance in Experiment 2, the negative Gf findings from Experiments 1 and 3 suggest either that this one positive result was a fluke, or that the Semantic task's predictive power is not robust to particular experimental or subject variables that are now unknown. Neither interpretation, however, calls for much optimism about the use of Semantic Conceptual Span in future individual-differences research. Although a semantic STM system may exist, given the compelling findings from experimental and neuropsychological research (e.g., Haarman & Usher, 2001; Martin & Freedman, 2001), the clustered Semantic Conceptual task does not seem to provide evidence for its importance as an individual-differences variable.

One could argue, however, that our regression analyses were biased against the Semantic task because our STM composites included word span, which probably taps both phonological and semantic storage and therefore limits any potentially unique contribution of the Semantic task to complex abilities. By this view, if only the nonword STM tasks were entered first into the regression equations, the Semantic task should predict incremental variance in cognitive ability. We therefore re-ran all our regression analyses by entering only the nonword STM span task(s) before the three Conceptual tasks, and found virtually the same patterns of results. The only change was that the Orthographic task predicted unique variance in both verbal ability and Gf in Experiment 1. As in our original analyses, then, the only place where the Semantic task provided unique predictive power was for Gf in Experiment 2.

As a second and final attempt to find some encouraging news for the Semantic task, we re-ran our regressions using scores from only the Conceptual tasks' trials that required recall of the final cluster of the list. These recency trials should provide the purest measure of STM capacity because they suffer no retroactive interference from intervening items, and indeed, prior studies using such cued-recall tasks have found that final-cluster recall correlates most strongly with cognitive ability (Cohen & Sandberg, 1977; Cantor et al., 1991; Haarmann et al., 2003). So, in each experiment, we regressed verbal ability and Gf

scores onto nonword span scores first (to account for any phonological contributions to ability), followed by the recency portions of the three conceptual tasks. Only one of the six analyses found the Semantic task to predict unique variance in ability (in Experiment 2, Semantic recency scores accounted for 7% of the variance in verbal ability beyond that accounted for the nonword STM and other Conceptual tasks). Moreover, the significant Semantic-task contribution to Gf that we had found in Experiment 2 disappeared when considering only recency scores. We thus conclude, again, that our findings do not provide much support for the criterion validity of the Semantic Conceptual task.

Our findings may be at odds with those of Haarmann et al. (2003) because we took their advice and used only a *clustered* version of Conceptual Span, which is less sensitive to the ability to actively cluster random lists. In 3 experiments, Haarmann et al. (2003) found *nonclustered* Conceptual Span to correlate strongly with comprehension and reasoning, and Haarmann et al. (2005) did so, too. (And, when Haarmann et al. (2003) reported a strong Conceptual Span-Gf correlation in a pilot study, it seems to have also been a nonclustered task.) To date, then, only one published experiment (Haarmann et al., 2003, Experiment 3) has tested whether the clustered version of Conceptual Span correlates with measures of complex cognition. Although the clustered task did predict substantial variance in two tests of comprehension and anomaly detection, its correlations were a bit weaker than those for the nonclustered task, and they were not contrasted to those for STM span tasks (nor to those for nonsemantic Conceptual tasks).

We acknowledge that our clustered tasks were not identical to the Haarmann et al. (2003) clustered conceptual task because we presented clusters of varying sizes within each list in unpredictable positions. It is therefore possible that a clustered Semantic task with 4 words in every cluster would have performed better than our task, but we see no reason why it should. Indeed, Saito (2006) has reported that a clustered Semantic task with predictable clusters of 4 words, and a similarly clustered Color task (as in our Experiment 1), were both poor predictors of sentence comprehension measures relative to nonclustered Conceptual tasks and WM span tasks.

Our generally disappointing findings regarding the Semantic task therefore suggest that Clustered Conceptual span provides little predictive power beyond that already tapped by phonological STM span tasks (or WM span tasks). It seems likely, then, that some aspects of clustering ability at encoding, retrieval, or both, is largely responsible for any incremental predictive utility of nonclustered Conceptual Span beyond conventional STM or WM spans. We would thus revise the advice provided by Haarmann et al. (2003), and suggest that any future research with Conceptual Span tasks directly contrast the validity of clustered versus nonclustered versions of the Semantic task, as well as contrasting the validity of Semantic Conceptual tasks to nonsemantic tasks like the ones we developed here.

AUTHOR NOTE

We are grateful to Alex Cereceres, Maria Cichetti, Andrew Dean, Wilfred Drath, Josh Jensen, Lindsay LaPlaca, Daniel McCord, Misty Nichols, Jessica Sherard, Pamela Shue, and Megan Tinker for their assistance in data collection. Correspondence concerning this article should be sent to M. J. Kane, Department of Psychology, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170 (e-mail: mjkane@uncg.edu).

REFERENCES

BADDELEY, A. D., GATHERCOLE, S., & PAPAGNO, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**, 158-173.

BADDELEY, A. D., LEWIS, V. J., & VALLAR, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology*, **36**, 233-252.

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.

BATTIG, W. F., & MONTAGUE, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, **80**, 1-46.

BERGER, F. R., GUPTA, W. B., BERGER, R. M., & SKINNER, J. (1990). *Air Force Officer Qualifying Test (AFOQT) form P: Test manual* (AFHRL-TR-89-56). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

BOUSFIELD, A. K., & BOUSFIELD, W. A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports*, **19**, 935-942.

BUNTING, M. F. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 183-196.

BURGESS, N., & HITCH, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, **106**, 551-581.

CANTOR, J., ENGLE, R. W., & HAMILTON, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence*, **15**, 229-246.

COHEN, R. L., & SANDBERG, T. (1977). Relation between intelligence and short-term memory. *Cognitive Psychology*, **9**, 534-554.

COHEN, R. L., & SANDBERG, T. (1980). Intelligence and short-term memory: A clandestine relationship. *Intelligence*, **4**, 319-331.

COLTHEART, M. (1981). *Unpublished MRC database user manual: Version 1*.

CONWAY, A. R. A., COWAN, N., BUNTING, M. F., THERRIAULT, D., & MINKOFF, S. (2002). A latent variable analysis of working memory capacity, short term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, **30**, 163-183.

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.

COWAN, N. (2005). *Working memory capacity*. New York: Psychology Press.

CROSSON, B., RAO, S. M., WOODLEY, S. J., ROSEN, A. C., BOBHOLZ, J. A., MAYER, A., ET AL. (1999). Mapping of semantic, phonological, and orthographic verbal working memory in normal adults with functional magnetic resonance imaging. *Neuropsychology*, **13**, 171-187.

DANEMAN, M., & MERIKLE, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, **3**, 422-433.

DEMPSTER, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin*, **89**, 63-100.

EKSTROM, R. B., FRENCH, J. W., HARMAN, M. H., & DERMEN, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

ENGLE, R. W., & KANE, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.) *The psychology of learning and motivation* (Vol. 44, pp. 145-199). New York: Academic Press.

ENGLE, R. W., NATIONS, J. K., & CANTOR, J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology*, **82**, 799-804.

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.

FARRELL, S., & LEWANDOWSKY, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, **9**, 59-79.

HAARMANN, H. J., ASHLING, G. E., DAVELAAR, E. J., & USHER, M. (2005). Age-related declines in context maintenance and semantic short-term memory. *Quarterly Journal of Experimental Psychology*, **58A**, 34-53.

HAARMANN, H. J., DAVELAAR, E. J., & USHER, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory & Language*, **48**, 320-345.

HAARMANN, H. J., & USHER, M. (2001). Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review*, **8**, 568-578.

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, J. H. (1991). *Fundamentals of item response theory*. New York: Sage.

HANTEN, G., & MARTIN, R. C. (2000). Contributions of phonological and semantic short-term memory to sentence processing: Evidence from two cases of closed-head injury in children. *Journal of Memory & Language*, **43**, 355-361.

HENSON, R. N. A. (1998). Short-term memory for serial order: The start–end model. *Cognitive Psychology*, **36**, 73-137.

HORN, J. L., DONALDSON, G., & ENGSTROM, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. *Research on Aging*, **3**, 33-84.

JACOBS, J. (1887). Experiments on "prehension." *Mind*, **12**, 75-79.

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.

KELLOGG, C. E., & MORTON, N. W. (1999). *Revised Beta Examination—Third edition*. San Antonio, TX: Psychological Corporation.

KLINE, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

LA POINTE, L. B., & ENGLE, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 1118-1133.

LUSTIG, C., MAY, C. P., & HASHER, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, **130**, 199-207.

MANDLER, G. (1968). Organized recall: Individual functions. *Psychonomic Science*, **13**, 230-236.

MARTIN, R. C., & FREEDMAN, M. (2001). Verbal working memory: The ins and outs of phonological and lexical-semantic retention. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 331-349). Washington, DC: American Psychological Association.

MARTIN, R. C., & ROMANI, C. (1994). Verbal working memory and sentence comprehension: A multiple-components view. *Neuropsychology*, **9**, 506-523.

NEATH, I., & NAIRNE, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, **2**, 429-441.

OBERAUER, K. (2005). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 393-407). Thousand Oaks, CA: Sage.

OBERAUER, K., SCHULZE, R., WILHELM, O., & SÜSS, H.-M. (2005). Working memory and intelligence—their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, **131**, 61-65.

PAGE, M. P. A., & NORRIS, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, **105**, 761-781.

PSYCHOLOGICAL CORPORATION (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Psychological Corporation.

RAVEN, J. C., RAVEN, J. E., & COURT, J. H. (1998). *Progressive matrices*. Oxford: Oxford Psychologists Press.

SAITO, S. (2006, July). *Forgetting and interference in working memory span tests*. Paper presented at the 4th International Conference on Memory, Sydney, Australia.

SWANSON, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, **84**, 473-488.

TEHAN, G., & LALOR, D. M. (2000). Individual differences in memory span: The contribution of rehearsal, access to lexical memory, and output speed. *Quarterly Journal of Experimental Psychology*, **53A**, 1012-1038.

TURNER, M. L., & ENGLE, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, **28**, 127-154.

NOTES

1. Swanson (1992) created two immediate memory tasks for use with children that closely resemble Conceptual Span, but the data for these tasks were collapsed with several other memory tasks and so we cannot determine how well they individually predicted the criteria measures of academic achievement.

2. We thank Henk Haarmann for providing us with the stimuli and programming codes for the Pronoun Text and Anomalies tasks.

(Manuscript received April 19, 2006;
revision accepted for publication August 3, 2006.)