Behavior Research Methods
2005, 37 (4), 581-590

Comparison of four scoring methods for the reading span test

NAOMI P. FRIEDMAN and AKIRA MIYAKE
University of Colorado, Boulder, Colorado

This study compared four common methods for scoring a popular working memory span task, Daneman and Carpenter's (1980) reading span test. More continuous measures, such as the total number of words recalled or the proportion of words per set averaged across all sets, were more normally distributed, had higher reliability, and had higher correlations with criterion measures (reading comprehension and Verbal SAT) than did traditional span scores that quantified the highest set size completed or the number of words in correct sets. Furthermore, creation of arbitrary groups (e.g., high-span and low-span groups) led to poor reliability and greatly reduced predictive power. It is recommended that researchers score span tasks with continuous measures and avoid post hoc dichotomization of working memory span groups.

Complex working memory span tasks, which require participants to fulfill both processing and storage requirements, are widely used in various areas of psychology (Miyake, 2001). Despite their popularity, no standard scoring methods exist; rather, researchers typically select a scoring method used in previous research. In some cases, researchers may try different scoring methods and select the one that seems the best, but the criteria used can vary widely across research labs. So far, little systematic effort has been made to compare and evaluate different scoring methods. In this report, we present in-depth comparisons of four common methods for scoring working memory span tasks, focusing on what is arguably the most prevalently used working memory task of all, Daneman and Carpenter's (1980) reading span test.

In the reading span test, participants read aloud sets of two to five or six sentences and attempt to remember the last word of each sentence. Usually, the participants begin with the easiest trials (those with two sentences) and work up to the most difficult ones (those with five or six sentences), but variations in which the trials are randomized also exist (e.g., Engle, Cantor, & Carullo, 1992; Friedman & Miyake, 2004b). In many cases, the task is terminated once a participant "fails" a level (e.g., if a participant fails to recall a majority of the trials in a level; Daneman & Carpenter, 1980), but some researchers prefer to administer all the trials to all the participants (e.g., Shah & Miyake, 1996). The latter procedure permits a wider range of scoring methods but has the disadvantage that the participants may become frustrated with the task once it advances beyond their abilities.

The traditional span score is the highest level at which the participant recalls a majority of the trials (e.g., two out of three sets, as was done by Daneman & Carpenter [1980], or three out of five sets, as was done by Miyake, Just, & Carpenter [1994]). In addition to the traditional span score, some researchers adopt other scoring methods, such as counting the total number of words in perfectly recalled sets (Engle, Tuholski, Laughlin, & Conway, 1999; McNamara & Scott, 2001), the total number of words recalled (Friedman & Miyake, 2000; Tirre & Peña, 1992; Turner & Engle, 1989), or the proportion of words per set averaged across all sets (Kane et al., 2004). Little has been said in the literature about different scoring methods, except to note that they often correlate highly and usually show the same patterns of results (e.g., Klein & Fiss, 1999; Turner & Engle, 1989; Waters & Caplan, 1996). However, even if different scoring methods are highly correlated, there may still be nonnegligible differences among them. Thus, in the present study, we compared and evaluated four scoring methods in terms of their distributional characteristics, reliability, and criterion validity.

The version of the reading span test used in this study had six trials at each of four levels (i.e., two to five sentences per set). The task was structured so that the participants completed three sets at each level and then completed another three sets at each level, for a total of six trials per level (i.e., a total of 84 target words to recall). In other words, each half of the reading span test was like a shorter version in which there were only three trials per level. This design allowed us to calculate internal reliability in two ways (i.e., split-half reliability and Cronbach's alpha) and also to examine the reliability of a version with fewer trials per level (i.e., three trials per level or the first half of the task). In addition, the participants performed the reading span test on two occasions, allowing us to calculate the test–retest reliability.

This research was partially supported by a National Science Foundation (NSF) Graduate Research Fellowship (to N.P.F.) and by a research grant from the National Institute of Mental Health (MH63207) and a training grant from the National Institute of Mental Health (MH01865). Correspondence concerning this article should be addressed to N. P. Friedman, Institute for Behavioral Genetics, University of Colorado, 447 UCB, Boulder, CO 80309 (e-mail: naomi.friedman@colorado.edu) or to A. Miyake, Department of Psychology, University of Colorado, 345 UCB, Boulder, CO 80309 (e-mail: miyake@colorado.edu).

The four scoring methods we examined were as follows.

1. *Total words*: the total number of words recalled across all trials. For example, if a participant recalled three out of five words on a trial, he or she received three points for that trial. Because this score included words recalled from a set even if the other words in that set were not recalled, it picked up differences between individuals who could recall some words from each set and individuals who forgot most of the words in the set (e.g., individuals who could recall four out of five words at Level 5 vs. individuals who could recall only one out of five words at Level 5). The maximum possible score was 84.

2. *Proportion words*: the average proportional recall for each trial. Specifically, the proportion of words recalled was calculated for each trial (e.g., if a participant recalled three out of five words on a trial, he or she scored .6 for that trial), and then the proportions for all the trials were averaged (i.e., all the trials were weighted equally). With this scoring system, forgetting words at early levels resulted in lower overall scores than did forgetting words at later levels (e.g., forgetting one word at Level 2 resulted in a score for that trial of 50%, whereas forgetting one word at Level 5 resulted in a score for that trial of 80%). The maximum possible score was 1.00.

3. *Correct sets words*: the total number of words recalled in perfectly recalled sets. For example, if the participant recalled all five words on a Level 5 trial, that participant would receive 5 points for that trial. However, if a participant failed to recall all the words from that trial correctly, he or she received 0 points for that trial, regardless of how many words were missed (e.g., 0 points were given even if four out of five words were correctly recalled at Level 5). This score is equivalent to the number of sets recalled perfectly, weighted by the number of words in each set. Because this score does not count partially recalled sets, it does not discriminate between individuals who could recall some words from each set and individuals who forgot most of the words in the set, as the total words score does. The maximum possible score was 84.

4. *Truncated span*: the highest level at which the participant recalled a majority of sets (four out of six). In addition, analogous to the original scoring method used by Daneman and Carpenter (1980), the participants were given half a point for getting three out of six or a quarter point for getting two out of six at the subsequent level (e.g., if a participant recalled four sets at Level 2, four sets at Level 3, and three sets at Level 4, that participant would receive a span score of 3.5). To receive credit for a level, the participants had to have passed all the previous levels. This score is equivalent to a span score for a span test that is terminated after the participant fails a level. The maximum possible score was 5.0.

Our expectation was that more continuous scoring methods would have better distribution and reliability characteristics because they provide more discrimination in terms of individual differences (Miyake, Emerson, & Friedman, 1999). Although the distribution characteristics (i.e., normality tests, skewness, and kurtosis) for span measures are rarely reported, there are a few studies that have reported reliability for various working memory span measures. In particular, Waters and Caplan (1996) found that the truncated span method of scoring the reading span test had a test–retest reliability of only .41 (N = 44). Moreover, when they used this scoring method to classify individuals as high, medium, or low span, 41% changed classification from Session 1 to Session 2, with equal numbers of scores increasing and decreasing. In contrast, Klein and Fiss (1999) used a total words method of scoring the operation span test (Turner & Engle, 1989) and found test–retest reliability estimates ranging from .67 to .81. Similarly, using a total words scoring method (specifically, the overall percentage of words recalled) on a reading span test, Tirre and Peña (1992) found an internal consistency of .95 and a test–retest reliability estimate of .73.

Although it is difficult to compare reliability across studies in which different span tasks and administration procedures have been used, the apparently higher reliability of the total words score over the truncated span score suggests that more continuous scoring methods may be preferable. In fact, in a recent study of 139 individuals 18 to over 80 years of age, Waters and Caplan (2003) noted that a scoring method that quantified the percentage of items recalled generally resulted in better test–retest reliability than did span scores that quantified the highest level passed for working memory span tasks, although they did find higher reliability for truncated span scores in this study (.73 to .76 for syntactically simple and complex versions of the reading span test, respectively) than in their previous study (Waters & Caplan, 1996). In the present study, we build on these findings from Waters and Caplan (2003) but go beyond their study by examining additional scoring methods (a total of four measures) and investigating the distributions and criterion validity, as well as the reliability, of these measures.

In the present study, we also examine what happens when span scores are used in a post hoc manner to assign individuals to discrete groups, using median splits (e.g., high-span and low-span groups). Such discrete groups are often created so that traditional statistical analyses (e.g., an ANOVA) can be used to examine how individuals with different working memory capacities differ in their performance on a criterion task (e.g., online processing of syntactically complex sentences), particularly when researchers are interested in testing the interactive effects between the span variable and other experimental variables of interest (e.g., the levels of syntactic complexity of sentences). From a statistical viewpoint, however, such group analyses can raise problems, because they treat all members in a group as identical despite variation in the actual span scores, hence reducing power (Humphreys & Fleishman, 1974; MacCallum, Zhang, Preacher, & Rucker, 2002; McClelland, 1997). According to Cohen (1983), dichotomizing a continuous variable can reduce its variance by as much as 20%–60%, with a loss of power equivalent to discarding one to two thirds of the sample. Given the ease of including continuous variables and their interactions in multiple regression analyses, such losses are unnecessary, but post hoc dichotomization of quantitative variables is still common practice (MacCallum et al., 2002), particularly in quasi-experimental studies of working memory that include participants' working memory span capacities as a key factor of interest. Hence, we will provide specific comparisons between continuous and dichotomized data in order to add another concrete example to the growing methodological literature documenting the statistical costs associated with this practice.

Table 1
Descriptive Statistics for Reading Comprehension Measures

Task	Mean	*SD*	Range	Skewness	Kurtosis
Reading Comprehension					
Session 1	18.98	4.25	6–23	0.09[*]	−0.63[†]
Session 2	13.02	3.85	3–22	−0.18[*]	0.22[†]
Verbal SAT	541.89	79.89	370–770	0.03[‡]	−0.11[§]

Note—*Mean* for sessions refers to number of items correct out of 25. None of the skewness and kurtosis values were significantly larger than zero. [*]Standard error = 0.26. [†]Standard error = 0.52. [‡]Standard error = 0.28. [§]Standard error = 0.55.

METHOD

Participants

The participants were 83 undergraduates at the University of Colorado at Boulder, who participated to partially fulfill a course requirement for an introductory psychology class. All were native English speakers. These participants were part of a larger study (Friedman & Miyake, 2004a) that included measures not described here, as well as a different administration condition for the reading span test. The participants included in the present analyses were those in the *experimenter-administered* condition, in which the participants were not allowed any time beyond that needed to read the sentences. This condition, which is close in spirit to the original reading span task (Daneman & Carpenter, 1980), was determined to be more valid as a predictor of complex cognition than was a *participant-administered* condition, in which the participants could take as much time as they wanted to read the sentences (for details, see Friedman & Miyake, 2004a). Data for 1 additional participant were excluded because he did not complete the second session.

Materials and Procedure

Each task had two parallel versions, created to assess test–retest reliability. The two versions were completed in two separate sessions.

Reading span measures. Two parallel forms of the reading span test (Daneman & Carpenter, 1980) were created by randomly assigning stimuli from a common set of stimulus materials to each task, so that the two tasks were matched in terms of the sentence lengths and difficulties, as well as the lengths of the to-be-recalled words. Each task began with two practice trials at Level 2, then proceeded through three trials at each level (2–5), and then three at each level again, for a total of six trials at each level (i.e., a total of 84 target words to be recalled).

This structure allowed for the calculation of internal reliability in two ways: split-half reliability and Cronbach's alpha. Furthermore, the administration of the reading span test in two halves (i.e., three trials at Level 2, Level 3, Level 4, and Level 5 and then three trials at each level again) allowed us to obtain a reading span score for a version with only three trials per level, so that we could examine the impact of administering a short version of the reading span test. When each half of the tasks was scored, the scoring procedures were the same as those described earlier, except for the truncated span score, which was calculated as the highest level at which the participant recalled at least two out of three sets, with an additional half a point for completing one set at the subsequent level (Daneman & Carpenter, 1980). In addition to the multiple estimates of internal reliability, we also calculated the test–retest reliability by correlating the span scores for the first session with those for the second session.

The procedure for the reading span tests was as follows. The experimenter pressed a button on a button box to bring up each sentence for the participant to read aloud. The participants were warned that the last word of each sentence would appear after a 1-sec delay, so they would not be tempted to look at it first. The experimenter pressed the button for each new sentence as soon as the participant finished pronouncing the last word of the current sentence, and the next sentence appeared 250 msec after the button had been pressed. The experimenter emphasized the importance of beginning to read each sentence aloud as soon as it appeared, and she reminded the participants of this requirement whenever she detected pausing. After each set, three red question marks appeared in the center of the screen, to signal the recall period, and remained there until the participant indicated that he or she was finished.

When orally recalling the sentence-final words, the participants were to try to recall them in order or, at least, not say the last word first. If a participant did begin recall with the last word, the experimenter counted it correct only if he or she could repeat (upon the experimenter's request) that word after he or she had finished recalling the other words. Text instructions on the computer signaled changes in levels.

Reading comprehension measures. The reading comprehension tests were modified versions of practice SAT tests published in Brownstein and Weiner (1974). Each test had eight passages, and each passage had two to five multiple-choice questions (five alternatives) that tested the participants' abilities to remember facts and make inferences about the material. In explaining the procedure, the experimenter stated that the tests were similar to SAT or ACT reading comprehension tests and that they should not guess, because they would lose a quarter of a point for each wrong answer (the actual scoring procedure, however, did not penalize for incorrect answers, because the penalization resulted in slightly less reliable measures). The participants had 20 min to finish as much of the test as possible. If they finished early, they were encouraged to check over their answers, but they were not required to do so if they wanted to begin the next task at that point. The reading comprehension score was calculated as the total number of questions answered correctly.

In addition, the participants provided written consent for the experimenter to obtain their scores for the Verbal sections of the SAT or ACT from official university records when they were available (nine were not). ACT scores for 12 participants were converted to SAT scores by averaging the ACT English and Reading component scores, converting this average to a *z* score, using an ACT mean of 20 and a standard deviation of 5, and then converting these *z* scores to SAT scores, using a SAT mean of 500 and a standard deviation of 100 (test norms taken from Glass & Hopkins, 1996). Table 1 provides descriptive statistics for the reading comprehension measures and the Verbal SAT scores.

General Procedure

Each participant was tested individually. In Session 1, the participant read and signed a consent form, then completed the reading span task and reading comprehension task, as well as other tasks not analyzed here. The participant returned for the second session approximately 2 weeks later and completed the parallel versions of the same tasks in the same order as that in Session 1.

RESULTS

For each scoring method, we examined distribution characteristics and reliability, as well as correlations with criterion validity measures (i.e., reading comprehension

Table 2
Correlations Among the Different Scoring Methods for the Reading Span

Scoring Method	1	2	3	4
Session 1				
1. Total words	–			
2. Proportion words	.99	–		
3. Correct sets words	.90	.89	–	
4. Truncated span	.81	.82	.89	–
Session 2				
1. Total words	–			
2. Proportion words	.99	–		
3. Correct sets words	.88	.88	–	
4. Truncated span	.81	.82	.89	–

Note—All correlations are within sessions. All *p*s < .05.

and Verbal SAT). In addition, we analyzed the data using median splits of the scores.

Distribution Characteristics

As is shown in Table 2, the four scoring methods were highly correlated in both sessions. The highest correlations were between total words and proportion words ($r = .99$ in both sessions), indicating that these two scoring methods were virtually identical. Although the other scoring methods were also highly correlated with these two (*r*s ranged from .81 to .90), there was some evidence that they were not equivalent in terms of distributions.

Table 3 presents descriptive statistics for the four scoring methods for each session, including skewness and kurtosis information. For the total words and proportion words measures, the skewness and kurtosis statistics were not significantly larger than zero (i.e., the 95% confidence interval calculated with the reported standard errors did not include zero). In contrast, the skewness and kurtosis values for the correct sets words and truncated span scores were significantly different from zero, suggesting that the distribution characteristics of these more discrete scoring methods are less desirable.

This conclusion is corroborated by normal quantile–quantile plots that we created for each measure for each session.[1] These graphs, shown in Figure 1, plot the quantiles of a variable's distribution against the quantiles of the normal distribution, enabling deviations from normality to be easily detected visually. For each graph, the vertical (y) axis plots the quantiles for the expected normal distribution, and the horizontal (x) axis plots the quantiles for the observed data. The solid line represents the expected value, given a normal distribution, and the circles represent the actual data points. Deviations of the data from the line are deviations from normality. In the lower right-hand corner of each plot, the Shapiro–Wilks *W* statistic for normality, which formally compares the quantiles of the fitted normal distribution with the quantiles of the data, is printed, along with the p value for $N = 83$. Lower values of this statistic (i.e., those with p values $< .05$) indicate significant deviance from normality. As is clear from the figure, the total words and proportion words scores tended to show the most normal distributions (i.e., the least deviation from the expected normal lines). In both sessions, these two scores had nonsignificant Shapiro–Wilks statistics. In contrast, the correct sets words and truncated span scores tended to be less normally distributed, with significant Shapiro–Wilks statistics.

Reliability

Reliability information for the four scoring methods is presented in Table 4A. For each method of scoring the span tasks, internal reliability was calculated in two different ways: (1) split-half (first half/second half) correlation adjusted with the Spearman–Brown prophesy formula and (2) Cronbach's alpha, which was calculated with the number correct summed across levels for the first trial at each level, the second trial at each level, and so on through the sixth trial at each level. Test–retest reliability was calculated as the Session 1–Session 2 correlation. As is shown in Table 4A, the total words, proportion words, and correct sets words scores showed reasonably high internal and test–retest reliabilities. The truncated span scores were less reliable, however, with the test–retest reliability dropping below the commonly used criterion of .70 (Nunnally, 1978).

The reliability estimates summarized in Table 4A were calculated using a version of the reading span with six trials at each level. Because reliability estimates, particularly Cronbach's alpha, increase with the number of trials, these estimates likely represent something close to the upper limit that might be obtained in other studies. This

Table 3
Descriptive Statistics for the Different Scoring Methods for the Reading Span

Scoring Method	Mean	*SD*	Range	Skewness†	Kurtosis‡
Session 1					
Total words	52.14	8.25	35–74	0.12	0.05
Proportion words	.68	.09	.46–.90	−0.13	−0.04
Correct sets words	19.95	8.87	6–53	1.30*	2.61*
Truncated span	2.35	0.48	1.5–4.25	1.52*	3.73*
Session 2					
Total words	52.20	8.46	33–78	0.46	0.46
Proportion words	.68	.09	.44–.94	0.19	0.21
Correct sets words	21.23	10.01	6–57	1.55*	2.87*
Truncated span	2.43	0.54	1.5–4.25	1.31*	2.24*

†Standard error = 0.26. ‡Standard error = 0.52. *$p < .05$.

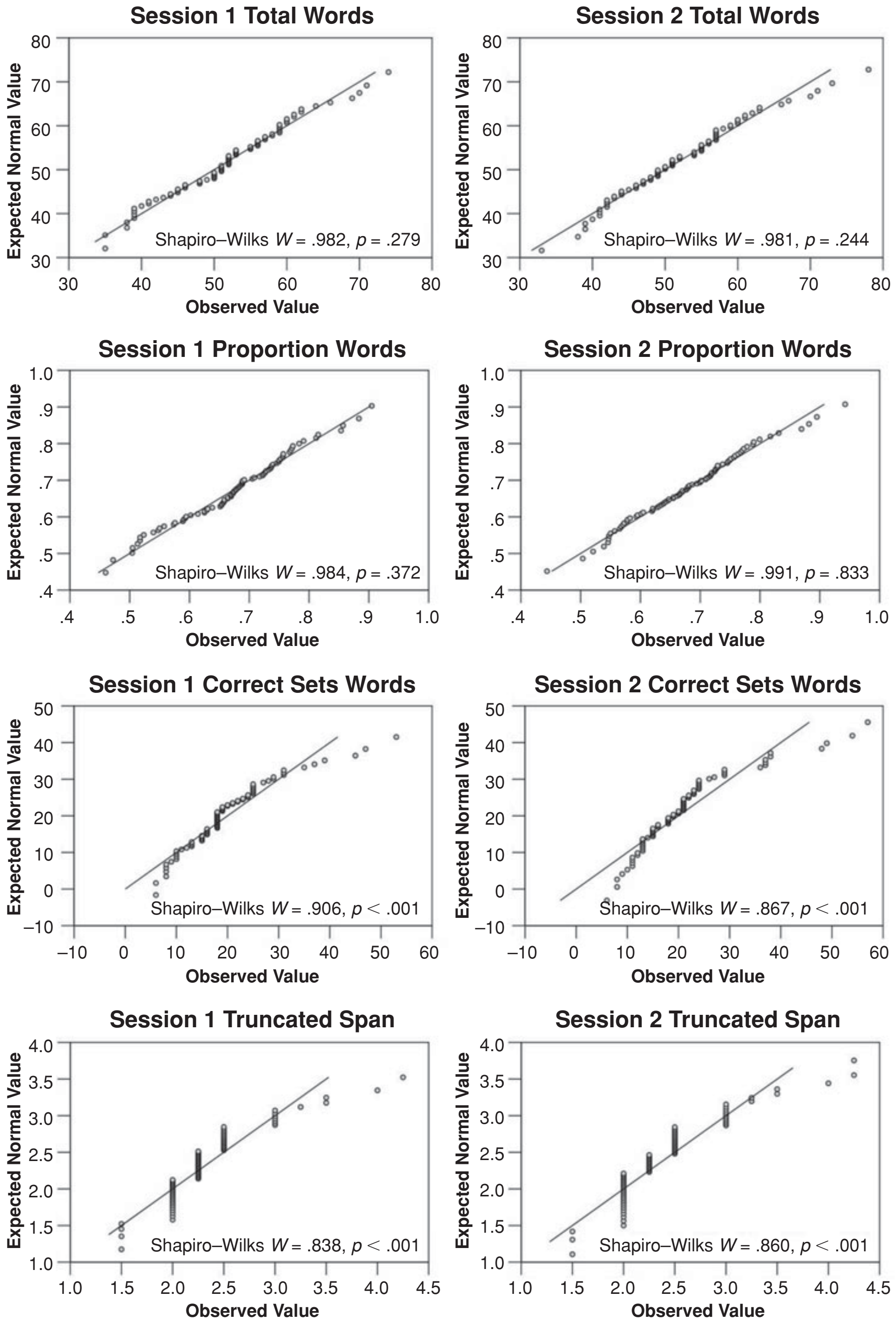

Figure 1. Normal quantile–quantile plots for the four scoring methods (Sessions 1 and 2) for the reading span task with six trials per level. For each graph, the solid line represents the expected value, given a normal distribution, and the circles represent the actual data points. Deviations of the data from the line are deviations from normality. In the lower right-hand corner of each plot, the Shapiro–Wilks *W* statistic for normality, which compares the quantiles of the fitted normal distribution with the quantiles of the data, is printed, along with the *p* value for $N = 83$. Lower values of this statistic (i.e., those with *p* values of <.05) indicate significant deviance from normality.

Table 4
Reliabilities for the Different Scoring Methods for the Reading Span

Scoring Method	Split Half* S1	Split Half* S2	Alpha† S1	Alpha† S2	Test–Retest
A. Six Trials Per Level					
Total words	.83	.92	.84	.86	.80
Proportion words	.83	.91	.84	.85	.79
Correct sets words	.80	.79	.73	.79	.74
Truncated span	.70	.75	–	–	.68
B. Three Trials Per Level					
Total words	–	–	.65	.68	.72
Proportion words	–	–	.64	.68	.72
Correct sets words	–	–	.39	.61	.59
Truncated span	–	–	–	–	.52

Note—S1, Session 1; S2, Session 2. *Split-half reliabilities calculated as correlation between first and second halves, adjusted with the Spearman–Brown prophesy formula. †Cronbach's alpha calculated with the number correct summed across levels for the first trial at each level, the second trial at each level, and so on through the last trial at each level. Alphas could not be calculated for the truncated span, because it was not possible to calculate a span score with only one trial per level.

means that, even though the reliability estimates for all of the scoring methods shown in Table 4A may seem acceptable, chances are that the situation would be considerably worse with more abbreviated versions of the task. The design of the reading span test in the present study enabled us to examine this possibility. Specifically, because the task began with three trials at each level and then cycled through the levels again, scores for the first half of the task were equivalent to what would be obtained if testing were terminated with only three trials per level, a length that is more typical in the literature (e.g., Daneman & Carpenter, 1980; Friedman & Miyake, 2004b).

The reliability estimates with only the first half of the reading span test in each session are presented in Table 4B. Because these analyses were based on only the first half of the span task in each session, it was not possible to compute split-half reliability. As is shown in Table 4B, none of the measures showed high internal reliabilities, although the total words and proportion words measures were closer to the recommended value of .70 (Nunnally, 1978) than was the correct sets words measure, which had unacceptably low reliability in the first session (.39). As for test–retest reliability, the total words and proportion words measures showed acceptable test–retest reliabilities (.72 for both), but the correct sets words and truncated span scores were less reliable (.59 and .52, respectively). These results indicate that the use of an abbreviated version of the reading span test (e.g., only three trials at each level) accentuates the reliability advantages of the total words and proportion words scores.

Correlations With Reading Comprehension

Table 5 presents the correlations of the four scores (calculated with six trials per level) with the two criterion measures. The reading span test has been known to correlate moderately with measures of global comprehension abilities. Daneman and Merikle (1996), for example, reported an average correlation of .41 (95% confidence interval = .38 to .44) in their meta-analysis of 77 studies. As Table 5 indicates (see the columns labeled "Raw"), the total words and proportion words scores showed the best correlations with both criterion measures, particularly in Session 2. The correlations with the correct sets words and especially the truncated span scores were lower.

To examine whether the correlations using the different scoring methods were significantly different from each other, we computed t values for each comparison, using the formula for comparing nonindependent correlations (Cohen & Cohen, 1983; Steiger, 1980). This formula takes into account the degree of correlation between the two measures compared: The higher the correlation between them (XY), the smaller the statistically detectable difference in correlation magnitude between XZ and YZ (where Z is the criterion measure). The results of comparisons using this formula indicated that in Session 1, truncated span ($r = .32$) was a significantly worse predictor of SAT Verbal than were total words [$r = .45$; $t(80) = 2.12$, $p = .037$] and proportion words [$r = .44$; $t(80) = 2.00$, $p = .049$]. In Session 2, correct sets words ($r = .42$) was a significantly worse predictor of reading comprehension than was proportion words [$r = .51$; $t(80) = 2.15$, $p = .035$], and truncated span ($r = .38$) was a significantly worse predictor of reading comprehension than were total words [$r = .51$; $t(80) = 2.19$, $p = .031$] and proportion words [$r = .52$; $t(80) = 2.45$, $p = .016$].[2] These results further highlight the superiority of the total words and proportion words measures over the correct sets words and, especially, truncated span measures.

Outliers

One possible explanation for the differences in the correlations with different scoring methods is that particular scoring methods may be more likely to yield outliers. To

Table 5
Correlations Between Reading Comprehension and the Different Scoring Methods for the Reading Span With and Without Outliers Removed

	Reading Comprehension*		Verbal SAT	
Scoring Method	Raw	Without Outliers	Raw	Without Outliers
Session 1				
Total words	.43	.53 (2)	.45	.47 (3)
Proportion words	.43	.52 (2)	.44	.46 (3)
Correct sets words	.45	.51 (8)	.40	.35 (5)
Truncated span	.37	.37 (7)	.32	.25 (4)
Session 2				
Total words	.51	.54 (5)	.47	.44 (3)
Proportion words	.52	.53 (4)	.46	.43 (3)
Correct sets words	.42	.50 (8)	.42	.37 (6)
Truncated span	.38	.37 (7)	.36	.18 (5)

Note—The numbers of outliers removed are given in parentheses. Outliers were defined as observations with levers greater than .10, Studentized $t > |2.00|$, or Cook's D values that were much larger than those for the rest of the observations. All $ps < .05$. *Correlations with Session 1 reading comprehension for Session 1 reading span and with Session 2 reading comprehension for Session 2 reading span.

examine this possibility, we computed the correlations with outliers removed. To identify outliers, we calculated for each correlation leverage, Studentized *t*, and Cook's *D* values, which assess how much influence a single observation has on the overall results. These statistics were selected because they are sensitive to different types of outliers (Judd & McClelland, 1989). The effects of removal for any participants with large values for these statistics (i.e., levers greater than .10, Studentized $t > |2.00|$, or Cook's *D* values that were much larger than those for the rest of the observations) were determined for each correlation.

The results are presented in Table 5 for both reading comprehension and Verbal SAT measures. The columns labeled "Without Outliers" list each correlation after all outliers fitting the above criteria were removed (the number of observations removed for each correlation is listed in parentheses). The comparison with the original, raw correlations gives an idea of how much the correlations changed with outliers removed. The largest changes were for Session 1 total words with reading comprehension (the correlation went up by .10 with two outliers removed) and for Session 2 truncated span with Verbal SAT (the correlation went down by .18 with five outliers removed).

Two conclusions are evident from Table 5. First, in each session, there were more outliers for the correct sets words and truncated span (four to eight outliers per correlation in Session 1 and five to eight outliers per correlation in Session 2) than for the total words or proportion words methods of scoring (two to three outliers per correlation in Session 1 and three to five outliers per correlation in Session 2). These results make sense, given the distribution problems associated with the correct sets words and truncated span measures, discussed earlier. Second, even with outliers removed, the truncated span method of scoring had lower correlations with reading comprehension and Verbal SAT than did the total words and proportion words scores.

Interestingly, removing outliers from the correlations with the total words and proportion words scores tended to increase the correlations, whereas removing outliers from the correlations with the truncated span scores tended to decrease the correlations, particularly for Verbal SAT. Examination of the scatterplots for each correlation revealed that when outliers showed up in the correlations with the total words and proportion words scores, they were mostly from individuals with aberrant scores on one of the measures (e.g., high reading span but low to average reading comprehension performance), instead of individuals with extreme scores on both of the measures. Because the former type of outlier has negative effects on correlations, removing those outliers tended to increase the correlation for the total words and proportion words measures. In contrast, both types of outliers were common for the correlations with the correct sets words and truncated span measures. In particular, the correlations with the Verbal SAT scores included some salient outliers of the latter type (i.e., those with distinctly high or low scores on both dimensions). Because these outliers can positively contribute to the overall magnitudes of correlations (assuming that they are not too deviant from the best-fitting regression lines), removing them had the effect of lowering the correlations with the Verbal SAT scores in the case of the correct sets words and truncated span measures. Thus, it appears that the predictive powers of the correct sets words and truncated span measures were more dependent on a small number of extreme observations that were either high or low on both dimensions and, thereby, nicely conformed to the expected pattern of positive correlations between reading span scores and the reading comprehension and Verbal SAT scores.

Impact of Post Hoc Arbitrary Grouping by Median Split

For each of the four scoring methods, we also examined the effects of a fairly common practice in the field, creating dichotomized groups by using median splits (i.e., all individuals with scores below the median are assigned to the *low-span* group, and those with scores above the median are assigned to the *high-span* group). In particular, we examined the loss of reliability and predictive power that might come with arbitrary post hoc dichotomization.

Regarding reliability, internal and test–retest reliability for span tasks using continuous scoring methods is typically satisfactory, as was shown earlier (Table 4A). When span scores are assigned to discrete groups, however, reliability can be much lower. For example, when Waters and Caplan (1996) assigned 44 participants to three reading span groups, using the truncated span scoring method, they found that 41% of the participants changed classification from Session 1 to Session 2. Waters and Caplan (2003) also found that assigning individuals to span groups on the basis of either their relative (e.g., top, middle, and bottom thirds of the distribution) or absolute (e.g., 2.5 or less for low spans, 4.0 or greater for high spans) scores resulted in poor stability (i.e., between 25% and 47% of the participants changed classification across sessions).

The present study corroborates these findings. Specifically, median-split classification changes from Session 1 to Session 2 were as follows: 23% of the participants for total words, 22% for proportion words, 31% for correct sets words, and 27% for truncated span. Hence, approximately one fourth of the 83 participants changed from high span to low span or vice versa from Session 1 to Session 2 (the two types of changes occurred approximately equally often). The stability of the median-split scores was fairly similar across different scoring methods, although the total words and proportion words split scores were slightly more stable. This result suggests that classification of participants into groups is problematic, regardless of what scoring methods are used to initially score the working memory measures.

We also examined the effect of post hoc dichotomization on the reading span's predictive power. Table 6 presents the percentages of variance accounted for in reading comprehension and Verbal SAT scores with the original unsplit scores and with split scores (i.e., only two values, one for individuals in the high-span group and zero for individuals in the low-span group). As Table 6 indicates, in all but one case, the median-split scores accounted for less variance in the criterion measures than did the continuous scores on which the median splits were based. On average, the median-split scores accounted for 7% less variance than did the original unsplit scores, and in several instances, the median-split scores accounted for 10%–15% less variance than the unsplit scores (e.g., total words scores in Session 2). These results confirm that post hoc dichotomization of a continuous variable into discrete groups is considerably less desirable than using the original unsplit scores, a problem that would likely be exacerbated when the goal of the study is to detect subtle, noncrossover interaction effects involving the span score variable (e.g., the effect of syntactic complexity would be more pronounced for low-span than for high-span participants).

Table 6
Percentage of Reading Comprehension Variance Accounted for by Regular Scores and Median-Split Scores for the Reading Span

	Reading Comprehension*		Verbal SAT	
Scoring Method	Rspan Unsplit	Rspan Split	Rspan Unsplit	Rspan Split
Session 1				
Total words	.19	.15	.20	.14
Proportion words	.18	.14	.19	.08
Correct sets words	.21	.08	.16	.04
Truncated span	.13	.14	.10	.03
Session 2				
Total words	.26	.15	.22	.07
Proportion words	.27	.17	.21	.13
Correct sets words	.17	.16	.18	.12
Truncated span	.14	.11	.13	.08

Note—Rspan, reading span; split, median split. *Variance accounted for in Session 1 reading comprehension with Session 1 reading span and in Session 2 reading comprehension with Session 2 reading span.

DISCUSSION

A comparison of four scoring methods for the reading span test led to clear and consistent results. Total words and proportion words scores had normal distributions, good reliability, and reasonably high correlations with reading comprehension measures. In contrast, correct sets words and truncated span scores fared less favorably on all of these attributes. The lower reliability for these two scoring methods was particularly evident when only three trials per level were used.

It is important to note that these differences in distributions, reliability, and criterion validity arose despite the fact that the correct sets words and truncated span scores were highly correlated (.81 to .90) with the total words and proportion words measures. Hence, high correlations do not necessarily guarantee that different scoring methods are practically equivalent and yield the same, equally preferable patterns of correlations with criterion measures. It may be the case that different scoring methods will result in qualitatively similar patterns of results, but it is likely that the results will be clearer when more continuous measures are used.

Another important finding was that when correct sets words and truncated span scores were used to predict reading comprehension and Verbal SAT scores, those correlations had more outliers than when total words or proportion words were used as predictors. When outliers were removed, the correlations with correct sets words and truncated span scores tended to decrease, whereas the correlations with total words or proportion words tended to increase or remain the same, suggesting that the predictive powers of the total words and proportion words measures are less critically dependent on certain types of extreme observations, as compared with correct sets words and truncated span scores.

The disadvantages of the correct sets words and, particularly, the truncated span scores make sense, given that they eliminate a good deal of individual-differences information (e.g., even though two individuals receive the same span score of 3.0, one could be much closer to 3.5 or 4.0 and the other to 2.0 or 2.5). When individual-differences information was further reduced by dichotomizing the scores, using median splits, the results were even worse. Approximately 25% of the participants switched from high to low span or vice versa across sessions. Moreover, the median-split scores accounted for 7% less variance, on average, than did the corresponding unsplit continuous measures. These problems are consistent with those reported by MacCallum et al. (2002). After examining and questioning the common reasons and justifications for post hoc dichotomization (e.g., the beliefs that it simplifies analyses or improves reliability, unawareness of its negative consequences, or a lack of familiarity with appropriate multiple regression analyses), MacCallum et al. concluded that “cases in which dichotomization is truly appropriate and beneficial are probably rare in psychological research” (p. 38).

It should be noted that splitting continuous variables into discrete groups for the purpose of preselecting participants for extreme group designs (e.g., administering a working memory span task to a large group of participants and selecting the top 25% and bottom 25% of the participants as high spans and low spans for later experiments) is not the same as dichotomizing continuous variables after data collection for the purpose of analyzing the data (for a recent example of extreme group designs, see Kane & Engle, 2003). Although extreme group designs can be susceptible to the *regression toward the mean* effect and may overestimate effect sizes, due to reduced error variance, they may have higher statistical power for detecting group differences and, particularly, interactions, because the standard errors associated with these effects tend to be smaller than those found with an unrestricted sample of scores (McClelland & Judd, 1993). Post hoc creation of arbitrary groups, however, leads to reduced power in almost all cases and sometimes can result in spurious differences between groups (MacCallum et al., 2002; Maxwell & Delaney, 1993).

In the present study, we examined only the reading span test in a sample of young college students. Thus, one must be cautious about generalizing the present results to the scoring of other working memory span tasks, such as the operation span test (Turner & Engle, 1989), as well as to other, cognitively more diverse samples, even though many of the findings we have reported in this article are consistent with various statistical principles delineated by methodologists. Despite this proviso, the results of the present study should serve as useful guidelines for scoring working memory span tasks and, possibly, even more traditional, storage-oriented short-term memory span tasks (e.g., digit or word span). Specifically, on the basis of the results reported here, we offer the following four general recommendations for scoring span tasks.

1. *Use the total words or proportion words scores and avoid the correct sets words and truncated span scores.* It is difficult to calculate the total words or proportion words scores if participants complete various numbers of trials. Thus, when these scoring methods are used, it is preferable to administer all the levels of the span task, rather than to stop administering the task once a participant fails at a particular level. As was mentioned in the introduction, one problem with administering all trials to all participants is that low-ability participants may become frustrated at the higher levels. One way to reduce this frustration is to randomize the order of the levels, so that the easy and the difficult trials are intermixed, rather than blocked. That way, low-ability participants will experience some success throughout the task, rather than just at the beginning. With respect to choosing between the total words and the proportion words scores, the results of this study indicate that these two methods produce essentially identical results. The total words score may be preferable in that it is easier to compute and conceptually more direct (i.e., it simply counts up the number of words recalled, with no weighting for the levels at which the words were recalled). However, it may also make sense to penalize more for the forgetting of words at easier levels, as the proportion words score does. It is clear from the results, however, that the correct sets words and truncated span scores should not be used, since both of them result in unnecessary data loss and poorer distributions.

2. *Administer at least three trials (preferably more) at each level of the span task to maximize the reliability of the span scores.* Longer working memory span tasks may place more of a burden on participants (and experimenters), so it makes sense to tailor the length of the span task for the needs of the study. If the purpose of the study is to obtain simple correlations with the working memory span tasks and many other measures are used, three trials per level may be adequate. However, if the study is designed to look for subtle differences in working memory abilities or interaction effects involving working memory abilities and/or if there is only one working memory measure in the study, using more trials per level is preferable.

3. *Regardless of which scoring method is chosen, examine the distributions for normality (or deviations from normality) and carefully evaluate the presence/absence of outliers and their impact on the pattern or magnitude of correlations.* Outliers may raise problems particularly in smaller correlational studies in which fewer participants are used. In the present study, a moderate sample size ($N = 83$) was used, and there were still several outliers, even with the total words and proportion words scores. In some cases, removing these outliers resulted in large changes in the correlations. With smaller samples, outliers would have even more influence on the results. Hence, it is important to make sure that the conclusions of the study do not rely on only a few extreme observations.

4. *Do not divide participants into discrete post hoc groups on the basis of their span scores*. Creating post hoc groups decreases power, because it treats all the individuals in each group the same, despite variation in their actual scores. Because measurement errors in one variable can be exacerbated when that variable is part of an interaction (McClelland & Judd, 1993), post hoc grouping variables can raise problems particularly when interactions are sought. Moreover, there is no good justification for using such post hoc grouping, because correlation or multiple regression techniques are more powerful when the effect of continuous span scores on a variable of interest is analyzed and are just as simple and flexible as ANOVAs in terms of adding interaction terms (see MacCallum et al., 2002, for a discussion and illustration of appropriate analyses).

REFERENCES

Brownstein, S. C., & Weiner, M. (1974). *Barron's how to prepare for college entrance examinations* (12th ed.). Woodbury, NY: Barron's Educational Series.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, **7**, 249-253.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

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

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, & Cognition*, **18**, 972-992.

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

Friedman, N. P., & Miyake, A. (2000). Differential roles for visuospatial and verbal working memory in situation model construction. *Journal of Experimental Psychology: General*, **129**, 61-83.

Friedman, N. P., & Miyake, A. (2004a). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory & Language*, **51**, 136-158.

Friedman, N. P., & Miyake, A. (2004b). The relations among inhibition and interference control functions: A latent variable analysis. *Journal of Experimental Psychology: General*, **133**, 101-135.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, **66**, 464-472.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach.* San Diego: Harcourt Brace Jovanovich.

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

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.

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, **7**, 19-40.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, **113**, 181-190.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, **2**, 3-19.

McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 393-411). New York: Cambridge University Press.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, **114**, 376-390.

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition*, **29**, 10-17.

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General*, **130**, 163-168.

Miyake, A., Emerson, M. J., & Friedman, N. P. (1999). Good interactions are hard to find. *Brain & Behavioral Sciences*, **22**, 108-109.

Miyake, A., Just, M. A., & Carpenter, P. A. (1994). Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory & Language*, **33**, 175-202.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, **125**, 4-27.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, **87**, 245-251.

Tirre, W. C., & Peña, C. M. (1992). Investigation of functional working memory in the reading span test. *Journal of Educational Psychology*, **84**, 462-472.

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

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology*, **49A**, 51-79.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, **35**, 550-564.

NOTES

1. Although skewness and kurtosis statistics are often used to specify the distributional characteristics of the measures of interest, they may not be particularly effective in detecting a type of nonnormal distribution called the *Cauchy distribution*, which can have highly negative effects on the results of least-squares analyses, such as ANOVAs and multiple regressions (McClelland, 2000). Normal quantile–quantile plots are effective ways to visually detect Cauchy distributions.

2. Several other comparisons were marginally significant: Session 1 truncated span versus correct sets words in predicting reading comprehension and Verbal SAT, Session 2 total words versus correct sets words in predicting reading comprehension, and Session 2 truncated span versus total words and proportion words in predicting Verbal SAT.

(Manuscript received June 21, 2004;
revision accepted for publication September 21, 2004.)