Behavior Research Methods, Instruments, & Computers
1999, 31 (3), 429-432

The reliability and stability of the Turner and Engle working memory task

KITTY KLEIN and WILLIAM H. FISS
North Carolina State University, Raleigh, North Carolina

The present study explored the psychometric properties of Turner and Engle's (1989) operation span task, a widely used measure of working memory capacity. We administered the task three times to 33 college students, using equivalent test materials. The interval between the first and second administrations was 3 weeks, with 6–7 weeks between the second and third administrations. Alpha coefficients were all .75 or more. Recall accuracy decreased as operation set size increased. Raw test–retest correlations ranged from .67 to .81, the corrected reliability was .88, and stability scores ranged from .76 to .92. Performance improved from the first to the second test. Relative to reported reliabilities of other tasks used to assess individual differences in working memory capacity, the operation span task appears to have several statistical advantages.

In recent years, individual differences in working memory (WM) capacity have received a great deal of attention from cognitive psychologists. WM capacity is considered to be a relatively stable individual difference that is related to performance on a variety of other cognitive tasks (Kyllonen & Christal, 1990). WM differences are also of interest to investigators exploring other basic cognitive processes such as the role of attention in storage and retrieval (Rosen & Engle, 1997).

There are a number of measures of WM capacity, all requiring participants to process and store information simultaneously. Individuals must read a sentence (see, e.g., Daneman & Carpenter, 1980), make a judgment about a sentence's acceptability (Waters & Caplan, 1996), or solve a simple arithmetic problem (La Pointe & Engle, 1990) while remembering a word across varying numbers of sentences or arithmetic operations. Scores on these complex WM span tasks are used as both continuous and categorical measures. Engle, Cantor, and Carullo's (1992) work testing the relationship between complex memory span scores and verbal Scholastic Aptitude Test (SAT) scores illustrates the correlational approach. Verbal fluency differences in participants categorized as high versus low span on the basis of their operation-span scores are an example of the second approach (Rosen & Engle, 1997). Both these methodologies, correlational and quasi-experimental, rest on a single administration of the task with the presumption that WM measures are reliable. There is some evidence regarding the internal consistency of WM measures, but little attention has been devoted to estimating test–retest reliability or stability. Although Turner and Engle (1989) found internal consistency to range between .74 and .81 among operations of the same set size across trials of their operation word span task, they did not report an overall estimate of internal consistency. Waters and Caplan (1996) have reported test–retest correlations of .41 for Daneman and Carpenter's (1980) reading span task and .66 for their own task. Reliabilities of this magnitude do not meet Nunnally's (1978) .7 criterion of minimum reliability adequacy. In addition, 41% of Waters and Caplan's participants classified as high or low span on the basis of the first testing were no longer in that category at retest, with 18% of the sample who were high or low at initial testing being classified in the other category altogether (low or high) when retested. Waters and Caplan wrote that these data "call into question the stability of working memory span scores based solely on sentence-final word recall measures" (p. 61). They also noted an additional problem with the Daneman–Carpenter task—the failure of recall accuracy to decrease with increasing load beyond span.

Given these concerns about complex WM measures, we decided to investigate the psychometric properties of one of the more popular WM measures, the Turner and Engle (1989) operation span test. We examined the internal consistency, reliability, and stability of the measure across three administrations, including the extent to which individuals who are categorized at a particular level (e.g., in the top third of a sample) retain their status across time. Internal consistency is a measure of the degree to which different test items—that is, sets containing different numbers of operations—measure the same variable. Internal consisten^y is evaluated using coefficient alpha (Cronbach, 1951). Test–retest reliability is the correlation between multiple testings. Stability is essentially the degree to which the individuals maintain the same rank order across intervals regardless of their level on the variable of interest, and is commonly taken to mean a constancy of level (Costa, McCrae, & Arenberg, 1980). We estimated stability using two techniques: the stability coefficient

We thank Adriel Boals, who programmed the working memory task, and Randall W. Engle, who provided many helpful comments during the preparation of this report. Correspondence should be addressed to K. Klein, Department of Psychology, Box 7801, North Carolina State University, Raleigh, NC 27695 (e-mail: kitty_klein@ncsu.edu).

formulae recommended by Heise (1969) and the extent to which individuals who are categorized at a particular level retain their status across time. Cutoff scores for defining individuals as high or low span are often established using upper and lower quartiles of the sample (see, e.g., Cantor & Engle, 1993; Rosen & Engle, 1997). Such a method of classification may be problematic if WM span scores are unstable or unreliable.

We also investigated the degree to which recall accuracy decreases as operation set size increases. As the number of operations increases, WM load increases. To the extent that the task discriminates among individuals, there should be a decrease in the percentage of correct responses as the number of operations increases.

METHOD

Participants

Thirty-three college students volunteered for the study to earn partial research credit for their introductory psychology course. They were tested three times. The interval between the first and second testings was 3 weeks; the duration between the second and third administrations was 6–7 weeks. Nunnally (1978) has recommended an interval of about 2 weeks between testings to study short-range fluctuations. We selected the longer interval to study the longer term stability of scores within the time constraints of the academic semester.

Materials and Procedure

The operation word span task (La Pointe & Engle, 1990) consists of a series of simple arithmetic operations with an answer followed by a one-syllable word—for example, "(9 × 1) − 9 = 1 back." Each operation and its accompanying word was presented simultaneously, one equation–word pair at a time on a computer screen. The participants' task was to respond verbally whether the answer following the equals sign was true or false and then to say the word that followed the operation. The experimenter then advanced the program to the next operation. After varying sets of two to seven of these operations, a question mark on the screen prompted participants to write down as many of the one-syllable words from the previous set as possible. In all, three sequences containing one set of each size were presented. The first experimental set was always a two-operation set, followed by a three-operation set with the order of the following sets determined randomly. Roughly half of the answers to the arithmetic operations were correct; half were incorrect. Presentation order of the sets was the same for all individuals. Participants were shown one practice problem and were told that they should try to remember as many words as they could, in any order, when prompted.

Three versions of the WM task were prepared, one for each administration. Each version contained a different set of arithmetic operations and words to be remembered equated for frequency. The operations and words were selected from the pools developed by Cantor and Engle (1993).

Scoring the WM task. Investigators have used a number of methods to score tasks that measure WM capacity (see Engle, Carullo, & Collins, 1991). Turner and Engle (1989) calculated two different scores: the set-size memory span and the total memory span. The set-size measure was the maximum size of the set in which the participant recalled the words correctly two out of the three times they were presented. The total memory measure was the sum of the number of correctly recalled words recalled in any order. Turner and Engle also eliminated data from participants who gave incorrect answers to more than 80% of the arithmetic equations. In this experiment, we calculated both the set-size and the total memory scores. However, we counted only words that were associated with correctly solved arithmetic problems, reasoning that incorrect answers to these simple equations suggest that the participant may have been concentrating on simply remembering the words without engaging in the processing task.

RESULTS

Inspection of the data revealed high rates of accuracy for the arithmetic operations portion of the task. Five participants correctly recalled a word paired with an incorrectly answered equation; the maximum number of such correct recall/incorrect equation solutions was two.

We first examined the relationship between the two methods used to score the operation word span task. The correlations between the set-size memory span and total memory span at Times 1, 2, and 3 were .89, .92, and .91, respectively. Following Turner and Engle (1989), the remaining analyses were conducted on the total memory span scores.

To estimate internal consistency, we computed the alpha coefficients for each administration of the WM task. Alpha coefficients are based on the average correlations of scores on any size operation set with the total WM scores. The results of the analysis are shown in Table 1. Given the high recall accuracy for two-operation

Table 1
Item-Total Correlations and Accuracy for Different Operation Set Sizes

	Correlations of Item With Total			Mean Accuracy Across
Operation set size	Time 1	Time 2	Time 3	3 Testings (Proportion)
Two	.143	.024	.025	.990
Three	.405	.396	.403	.919
Four	.620	.710	.574	.753
Five	.527	.694	.763	.629
Six	.656	.665	.752	.556
Seven	.654	.639	.822	.460
Alpha	.749	.777	.752	
Alpha adjusted*	.776	.810	.829	

*Alpha adjusted is with Set Size 2 omitted.

sets, correlations of these items and total span score are low. The adjusted alpha was calculated for operation sets of sizes three to seven, omitting the two-operation sets.

We next examined the test–retest statistics for the WM span task. Table 2 displays the raw correlations between tests and the means for each test.

An analysis of variance indicated that there were differences in mean levels across the experiment [$F(2,64) = 9.75$, $p < .0002$]. Total memory span scores increased more than three items from Time 1 to Time 2, presumably evidence of practice effects; there was no further change between the second and third administrations.

Raw correlations across intervals were of roughly the same magnitude as the internal consistency measures. The lowest test–retest correlation (.66) was between the Time 1 and Time 3 administrations. Test–retest correlations did not differ as a function of including or excluding the 2-operation sets.

Heise (1969) demonstrated that simple test–retest correlations may not measure true reliability of a measure because it is affected by temporal instability in the variable as well as by errors of measurement. Heise derived a formula for analyzing test–retest correlations when a variable is measured on at least three occasions that allows the separation of true score instability from measurement error and thus provides an estimate of the true reliability of the measures. Using Heise's formula, the true reliability for the version of the Turner–Engle WM task we employed was .883.

For our first stability index, we used another formula Heise (1969) developed to obtain true score stability coefficients for a given interval. The stability coefficient for the interval between the first and second administrations was .821; between the second and third, .92; and between the first and third, .756. Even the lowest of these coefficients exceeds two of the three raw test–retest correlations. The magnitude of these stability coefficients suggests that although there may be differences in mean levels, particularly from the first to the second administration, individuals do not change differentially across administrations.

In a second analysis of stability, we examined whether individuals classified as high or low span would maintain this categorization across the testing intervals. Owing to the small sample size, we could not utilize strict definitions of quartiles. We characterized individuals in the approximately top quartile at each administration as high span; participants in the approximately bottom quartile were considered low span. On the first testing, participants with scores of 54 or more were classified as high span (n = 10); participants with scores of 45 or lower (n = 10) were identified as low span. One individual in each group failed to maintain the same classification on the second testing, yielding a classification error rate of 10%. A lower error rate (5%) occurred for Time 2–Time 3 stability. None of the Time 2 high-span participants (n = 10, WM scores exceeding 56) were classified as low span on the basis of their Time 3 scores. One person who was classified as low span on Test 2 (n = 9, WM scores below 49) was in the high-span group at Time 3 (WM scores above 56).

Table 2
Raw Test–Retest Correlations and Means for the Three Administrations

	Time 1	Time 2	Time 3
Time 1		.725	.667
Time 2			.812
M	50.00	53.12	53.88
SD	6.99	7.25	7.74
Range	36–66	41–70	42–74

The second goal of the study was to examine recall accuracy for operation sets of each size. Table 1 presents the percentage of words accurately recalled for each set size, averaged across the three administrations. As would be expected if increasing the number of operations increases working memory load and if individuals differ in their ability to recall under load, recall accuracy decreased as set size increased [$F(5,160) = 67.48$, $p < .0.0001$].

DISCUSSION

Taken together, the WM scores obtained from three administrations of the Turner and Engle (1989) WM task indicate that it is an extremely reliable measure. The internal consistency statistics were acceptable, with alpha coefficients averaging .75. Raw score test–retest correlations were high, and the overall corrected reliability, .88, indicates the task is reliable at least across the 10 weeks of this study. Stability coefficients were acceptable and were highest for the longer interval (6–7 weeks) between testings. Although the stability of participants' classification as high or low span was much higher than Waters and Caplan (1996) have reported, this index was the least satisfactory of those studied. There are two obvious reasons for the classification errors across administrations. First, with such a small sample size, the misclassification of even a single individual appears as a relatively large error in percentage terms. Second, the distinction between extreme groups in terms of scores is relatively small. At Times 1 and 2, the difference between high- and low-span groups was 7 points; at Time 2 it was 8 points. Given that these differences are roughly equivalent to the standard deviation of the measures, it is not surprising that the classification of a few individuals changed across administrations.

Recall accuracy declined as the number of operations required before recall increased. The extremely high recall accuracy and low variability for the two-set operations suggest that omitting these sets from the test would not alter its reliability in a college population.

One unexpected finding was the increase in measured WM span from the first to the second administration. It would appear that practice effects can account for this increase. The stability coefficient for this interval indicates that participants at all levels of capacity benefited similarly from their initial exposure to the procedure.

We should note that the high reliabilities we observed may not apply to other versions of the Turner and Engle (1989) WM task. In this research, we followed the procedure of Cantor and Engle (1993), in which participants essentially control the rate of presentation. Whether similar results would obtain if operations were presented for a fixed interval (see, e.g., Singer, Andrusiak, Reisdorf, & Black, 1992) is unknown.

An important question not answered by our data is why the psychometrics for the Turner and Engle (1989) operation word span task are generally superior to those Waters and Caplan (1996) reported for two other complex WM tasks, Daneman and Carpenter's (1980) reading span task and Waters and Caplan's sentence span measures. Although our internal consistency measures are similar, we observed markedly higher test–retest reliabilities. One reason for this inconsistency may be the varying retest intervals Waters and Caplan employed. They retested participants a single time, at intervals between 31 and 176 days later. Their relatively low test–retest correlations may reflect the benefits of practice for individuals tested after relatively brief intervals and the absence of such effects for those tested 6 months later. Alternatively, there may be some feature of the WM measures they examined that are particularly sensitive to transient influences.

In conclusion, our data do not support Waters and Caplan's (1996) view that complex measures of WM span are unreliable. The Turner and Engle task (1989) as employed in this study demonstrated high levels of internal consistency, test–retest reliability, and stability. Whereas one would anticipate high test–retest reliability across shorter time intervals, the high levels of reliability and stability across a 7-week interval suggest that researchers can have some confidence that the operation word span task reliably assesses individual differences in WM capacity.

REFERENCES

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1101-1114.

Costa, P. R., McCrae, R. R., & Arenberg, D. (1980). Enduring dispositions in adult males. *Journal of Personality & Social Psychology*, **38**, 793-800.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.

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

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., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research*, **84**, 253-262.

Heise, D. R. (1969). Separating reliability and stability in test–retest correlation. *American Sociological Review*, **34**, 93-101.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity! *Intelligence*, **14**, 389-433.

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.

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

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, **126**, 211-227.

Singer, M., Andrusiak, P., Reisdorf, P., & Black, N. (1992). Individual differences in bridging inference processes. *Memory & Cognition*, **20**, 539-548.

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**, 51-74.

(Manuscript received April 16, 1998;
revision accepted for publication August 31, 1998.)