APHASIOLOGY, 1998, VOL. 12, NO. 7/8, 771–785

Reliability and validity of an auditory working memory measure: data from elderly and right-hemisphere damaged adults

MARGARET T. LEHMAN
and CONNIE A. TOMPKINS

Department of Communication Science and Disorders, University of Pittsburgh, PA, USA

Abstract

The use of non-standardized measures in research and clinical assessments creates difficulties with interpretation and generalization of results obtained. One example of a widely used non-standardized tool is the reading/listening span paradigm for assessment of working memory (WM). WM is an important construct because of its purported relationship to language comprehension and capacity theories of cognition. This paper investigates several facets of reliability and validity for an auditory working memory measure designed for older adults and individuals with right hemisphere brain damage (RHD). Results from 28 non-brain-damaged subjects (NBD) and 11 RHD subjects indicate that the measure is internally consistent and reliable over time. Construct validity evidence, which compares favourably with evidence from existing literature, suggests that for NBD sujects this tool differentiates WM from simple short term memory. RHD subjects do not demonstrate the same pattern of validity results as the NBD group. Further evaluation with RHD patients is warranted, because clinically this tool may be useful as a measure of severity or a prognostic indicator of language comprehension abilities for this population.

Introduction

Non-standardized measures are widely used for research and clinical assessment. They are selected over standardized measures for various reasons, including ease of administration and shorter time requirements. Because non-standardized tasks can be tailored to specific needs of a situation, they may be more sensitive for a particular purpose. However, the lack of reliability and validity data for such tools limits the confidence with which conclusions can be made from the data obtained and prevents clear generalization of results beyond the immediate context.

One example of a non-standardized tool which has been used extensively is the reading/listening span task for investigating working memory (WM), or concurrent memory processing and storage. The construct of WM, as measured by this task paradigm, has been linked to several communicative abilities, such as reading comprehension (Daneman and Carpenter 1980, Waters *et al.* 1987),

Address correspondence to: Margaret T. Lehman, Department of Communication Science and Disorders, 4033 Forbes Tower, University of Pittsburgh, Pittsburgh, PA 15260, USA.

memory for discourse (Light and Anderson 1985), and inference revision (Daneman and Carpenter 1983, Tompkins *et al.* 1994). Working memory also is an important concept in capacity theories of language comprehension. The most prominent of these theories has been put forth by Just and Carpenter (1992) and revolves around the differences in comprehension skill of subjects with high and low working memory spans. Various modifications of Daneman and Carpenter's (1980) original tasks have been used to evaluate WM in a variety of populations, including young and older adults (Light and Anderson 1985, Salthouse and Babcock 1991) and patients with neurological disorders (Tompkins *et al.* 1994, Waters *et al.* 1995). Despite the widespread use of WM tasks, their reliability and validity are poorly documented and can only weakly be inferred from published reports.

This study evaluates several facets of reliability and validity for an auditory WM task designed for use with older adults and stroke patients (Tompkins *et al.* 1994). Tompkins and colleagues have reported preliminary data regarding the internal consistency and test–retest reliability of the measure, with promising results. Predictive validity evidence suggests that the measure has potential as a prognostic indicator, or an index of severity for adults with right hemisphere brain damage (RHD; Tompkins *et al.* 1994). This study expanded upon previous reliability data and provides a preliminary assessment of construct validity.

Methods

Subjects

Thirty-nine adults between the ages of 45 and 78 years participated in this study. Twenty-eight of the subjects were normally ageing adults with no known history of neurological damage or cognitively deteriorating condition. Eleven subjects had unilateral RHD due to cerebrovascular accident (seven thromboembolic, four haemorrhagic). CT or MRI scan reports were obtained for each RHD subject to verify that lesions were restricted to the right cerebral hemisphere. Post-stroke subjects had participated in a larger study of comprehension abilities following stroke conducted by the second author. Group data are provided in table 1. Clinical characteristics of the RHD group, which should facilitate generalization of reported results, are provided in table 2. These data indicate that the current RHD subjects are similar to RHD groups from previous research studies in terms of performance on auditory and visual comprehension and memory tasks (Tompkins 1990, 1991, Tompkins *et al.* 1992, 1994).[1]

All subjects reported completing at least 8 years of formal education and indicated negative history of alcohol or drug abuse. Handedness was determined by questioning potential subjects on the six most discriminating items from the Annet (1970) inventory; only those subjects who reported performing all actions with the right hand exclusively were included in the current study. All subjects were native speakers of English only. This was judged by enquiring whether they had learned any language besides English during early childhood. Additionally, all subjects passed a pure-tone hearing screening (35 dB at 500, 1000, and 2000 Hz).

[1] Because the subjects in the current study were not selected for exhibiting communication disorders, they may not be representative of RHD patients treated in a clinical setting. However, it is important to evaluate subjects with mild deficits as well as those who evidence more obvious disorders.

Table 1. Characteristics of two subject groups

	NBD (n = 28)	RHD (n = 11)
Gender	11 males 17 females	11 males
Age in years		
M (SD)	62·2 (7·6)	58·5 (9·2)
Range	47–74	44–73
Education*		
M (SD)	14·8 (2·3)	12·3 (2·9)
Range	12–20	8–16
No of days between tests		
M (SD)	22·5 (2·5)	21·4 (2·5)
Range	18–28	24–30
Estimated IQ[a]		
M (SD)	111·5 (6·5)	105·2 (8·4)
Range	99·3–112·0	92·4–116·0

* Significantly different.
[a] Wilson *et al.* (1979)

Table 2. Clinical characteristics of RHD subject group (n = 11)

	Mean (SD)	Range
Months post-onset	48·7 (32·2)	3–105
MMSE[a]	27·8 (1·5)	24–30
Neglect[b]	133·9 (21·9)	69–145
Receptive vocabulary[c]	156·9 (11·9)	134–171
Auditory comprehension		
BDAE[d]	92·9 (3·1)	88–98
Discourse comprehension test[e]	5·1 (1·8)	3–10
Tonal memory[f]	5·9 (2·5)	1–10
Judgement of line orientation[g]	22·5 (4·5)	15–29
Immediate story recall[h]	13·5 (1·9)	11–17

[a] Mini-Mental State Examination (Folstein *et al.* 1975).
[b] *Behavioural Inattention Test* (Wilson *et al.* 1987) (six conventional subtests only, maximum = 146, neglect cut-off score = 129).
[c] *Peabody Picture Vocabulary Test* (Dunn and Dunn 1981) (maximum = 175).
[d] Overall BDAE percentile (Goodglass and Kaplan 1983) (average of percentiles on the test's four auditory comprehension subtests).
[e] *Discourse Comprehension Test* (Brookshire and Nicholas 1993) (average error score for four stories, maximum = 16).
[f] Seashore *et al.* (1960) (three-tone level, maximum = 10).
[g] Benton *et al.* (1983) (corrected score, maximum = 35, normative data: 'defective performance' cut-off score = 18, 46% RHD normative sample scored 18 or below).
[h] *Arizona Battery for Communication Disorders of Dementia* (Bayles and Tomoeda 1990) (maximum = 17).

Finally, cognitive abilities were screened with the Mini-Mental State Examination (MMSE; Folstein *et al.* 1975). All NBD subjects had to score at least 27/30 to be included in the study.

Tasks

Working memory

Working memory is conceptualized as the concurrent processing and storage of information (Daneman and Carpenter 1980), in contrast to short-term memory (STM), traditionally defined as a passive storage buffer. The working memory measure evaluated in this study is a modification of Daneman and Carpenter's (1980) listening span task. Comprehensive details regarding construction of stimuli and administration protocol were summarized by Tompkins *et al.* (1994).

Stimuli were simplified for use with older adults and stroke patients and consisted of simple, active declarative sentences 3–5 words in length ($M = 4{\cdot}03$). The items most closely resemble the stimuli for Daneman and Blennerhassett's (1984) pre-school reading span task. The sentences were constructed to reflect common world knowledge, eliminating the historical facts included in Daneman and Carpenter's original stimuli; this was done so that task performance would hinge on linguistic processing and storage rather than academic knowledge.

One half of the sentences were true statements, the other half false. The truth of a statement could not be determined until the final word of the sentence, so subjects had to process the entire sentence before making a response. Sentence final words were moderate- to high-frequency (occurrence greater than 1 per million; Kucera and Francis 1967), 1–2 syllable common lexical items. Stimuli were recorded at a slow normal speech rate on to a high quality audiocassette tape. Duration of sentences was similar across items.

Stimulus validation was conducted with a group of normally ageing adults ($n = 11$) with no history of neurological insult or disease (Tompkins *et al.* 1994). To ensure a homogenous level of difficulty across stimulus sentences, validation subjects' reaction times for true/false judgements were measured. Any item that resulted in reaction times greater than 600 ms was altered, either by re-phrasing or re-recording, or was eliminated from the stimulus list (all subjects demonstrated 100 % accuracy on the judgements). A total of 42 sentences were thus created and validated. These were then divided into three sets each of 2–5 sentences. True and false statements were balanced within each level (see Appendix).

Measures of validity

Construct validity was assessed using two digit recall tasks (backward and forward recall) and a simple word recall task. The backward digit recall task from the *Wechsler Memory Scale—Revised* (WMS-R; Wechsler 1987) was chosen as a measure of convergent validity, because it requires subjects temporarily to store a string of digits while simultaneously 'juggling them around mentally' (Lezak 1995, p. 367), or reversing the presentation order for recall. Although the processing demands of the digit recall task may not be exactly the same as those required for language comprehension, backward digit recall is presumed to tap WM abilities due to the dual demands of the task (Lezak 1995, Salthouse 1990). Because the backward digit recall is a standardized task, it is perhaps the best available measure of dual processing and storage described in the literature.

Divergent validity was addressed with two different tasks. The first one was the forward digit recall task from the WMS-R. Forward digit recall is thought to solicit only STM abilities and thus should not measure the same construct as does the WM task (Daneman and Carpenter 1980, Tompkins *et al.* 1994, Turner and Engle 1989). This particular forward digit recall task was selected because it is a standardized measure. The second task for evaluating divergent validity was a simple word recall task. This is thought to assess STM because it requires only storage of the stimulus targets (Daneman and Carpenter 1980, Tompkins *et al.* 1994, Waters and Caplan 1996). In this task, subjects heard a lists of words (ranging from 3 to 7 items) and were asked to repeat the words. Although this task is more similar to the WM task than is forward digit recall in that subjects are required to recall a set of words rather than a closed set of digits, the lack of an additional processing requirement should differentiate it from WM.

Lexical items for the simple word recall task were selected to be similar to the sentence-final words of the working memory measure. Items were chosen from a list of high-frequency (greater than 80 per million; Kucera and Francis 1967) 1–2 syllable words. Familiarity ratings for lexical items were obtained from 10 normally ageing adults (similar in age, gender, and educational level to subjects in the current study). Questionnaires containing words representing a range of frequency of occurrence and predicted familiarity were mailed to the subjects. Subjects were instructed to rate each word on a scale of 1 ('not at all familiar') to 7 ('highly familiar'). Seventy-five items from this list that were high-frequency items (Kucera and Francis 1967) and were rated as 'familiar' or 'highly familiar' by all individuals who completed the questionnaires were selected for the word recall task. None of the items selected for the word recall task appeared in the WM sentences.

Words were arranged in three sets each of 3–7 items. Sets were constructed semi-randomly, with the following conditions: one two-syllable word appeared in each set; no set contained words that had obvious semantic relationships; and no two consecutive words formed a compound word (e.g. 'brush' could not follow 'hair').

Stimulus tapes for the digit and word span tasks were generated in a manner similar to those for the WM task (Tompkins *et al.* 1994). Stimuli were recorded in a sound-attenuated booth using a professional quality audiocassette recorder (Marantz PMD420) and high-quality audiocassette tapes (TDK metal type IV). All items were recorded with neutral intonation. Quality of the recordings was judged by two listeners for consistency of intonation across items and clarity of articulation and pronunciation. Stimuli were then digitized using the Creative Lab Sound Blaster (v. 2) sound-editing program. A 750 ms pause was inserted between digits or words within a set. A 3 s pause was inserted between sets. For the word recall stimuli, the carrier phrase 'set (x), ready?' was recorded at the beginning of each set. This alerting phrase was identical to that used in the WM measure. All stimuli were then transferred back to audiotape.

Equipment and procedures

Stimulus tapes were played from a high-quality tape recorder (Marantz PMD420) and presented binaurally to subjects through supra-aural headphones (Fostex T20). Volume was set to a comfortable level for each individual subject and maintained at that level throughout the testing session.

The two testing sessions were scheduled approximately 3 weeks apart (range

18–30 days). The first WM administration was completed in the first session, along with the hearing screening, MMSE, and other tasks related to a different, concurrently conducted study. In the second session, all memory tests were administered in the following order: digit recall, word recall, digit recall, WM. The order of forward and backward digit recall tasks was counterbalanced across subjects. RHD subjects also completed the six conventional subtests of the *Behavioural Inattention Test* (BIT; Wilson *et al.* 1987), a measure of visuo-spatial neglect.

Scoring

Working memory task

A combined error score was obtained by adding the errors from both recall and true/false components of the WM measure. This combined score was used in the correlational analyses, because by assessing accuracy of both the processing and memory components it provides a complete picture of working memory capacity. The combined score also takes into account trade-offs that appeared to occur for several subjects: When they recalled more words in a lengthy set, this increase came at the expense of true/false accuracy.

Digit recall tasks

For the digit recall tasks, digits had to be recalled in the order in which they were presented (or the opposite order for backward digit recall). Error scores were used instead of the traditional span scores, to circumvent the problems with a restricted range of scores which occurs with the span scores.[2] Error scores reflected the total number of digit transpositions or omissions.

Simple word recall task

Error scores consisted of the total number of omissions, incorrectly recalled targets (e.g. items from a previous set), and phonological/ perceptual errors (e.g. 'birth' for 'earth').

Results

Preliminary analyses

Group *t*-test results indicated no significant difference in performance on any of the experimental measures for males and females (all $t < 0{\cdot}95$, $p > 0{\cdot}10$) for the NBD subjects. Performance on the WM task, as measured by the total error score, also was not correlated to age, years of education, estimated IQ, or number of days between testing sessions for either subject group. WM error scores were not meaningfully related to gender for the NBD group (both $r < 0{\cdot}33$, $p > 0{\cdot}05$), or to severity of neglect (BIT score) for the RHD group (time 1: $r = -0{\cdot}26$, time 2:

[2] For purposes of comparisons with other published studies, analyses were conducted with span scores as well as error scores. The results obtained for the span scores were similar to those for the error scores, thus only the error score results will be reported here.

Table 3. Error score correlations (Spearman ρ)[a]

		Convergent	Divergent	
	WM2	Backward digit	Forward digit	Word recall
NBD group				
WM1	0·77*	0·69*	0·52*	0·59*
WM2		0·68*	0·77*	0·68*
RHD group				
WM1	0·77*	0·20	0·77*	0·97*
WM2		0·06	0·85*	0·74*

WM1 = WM combined error score time 1, WM2 = WM combined error score time 2.

[a] ρ values that exceed 0·51 for NBD subjects or 0·80 for RHD subjects are significantly different from 0 ($p < 0{\cdot}01$).

* Correlations meeting Cohen and Cohen's (1983) rule of thumb for a large effect in the behavioural sciences.

$r = -0{\cdot}42$, $p > 0{\cdot}05$).[3] Accordingly, these variables were not considered in further analyses.

Group differences over time were examined by comparing average WM combined scores across the two administrations. Average performance was not significantly different between time 1 and time 2 for either NBD ($M_1 = 5{\cdot}3$, $M_2 = 5{\cdot}4$, $t = -0{\cdot}32$, $p > 0{\cdot}10$, SEM = 0·56) or RHD groups ($M_1 = 9{\cdot}8$, $M_2 = 10{\cdot}1$, $t = -0{\cdot}30$, $p > 0{\cdot}10$, SEM = 0·90).

To assess the internal consistency of the other measures as administered in this study, split-half reliability coefficients with the Spearman–Brown adjustment were calculated. Results for NBD subjects indicate high internal consistency for backward digit recall ($r_{nn} = 0{\cdot}79$), forward digit recall ($r_{nn} = 0{\cdot}86$), and word recall ($r_{nn} = 0{\cdot}93$). For RHD subjects, internal consistency was not as high for the backward digit task ($r_{nn} = 0{\cdot}62$); however, it was strong for both forward digit recall ($r_{nn} = 0{\cdot}86$) and simple word recall ($r_{nn} = 0{\cdot}85$).

Reliability

Test–retest reliability

To supplement the analyses of differences over time, Spearman correlations were conducted. A moderately strong correlation was obtained for both groups (both $\rho = 0{\cdot}77$).

Internal consistency

Internal consistency of the WM measure was evaluated using a split-half reliability analysis. The test was divided into odd versus even items and recall accuracy scores on the two halves were correlated, using a Spearman–Brown adjustment to obtain

[3] It was predicted that WM performance may be correlated to severity of neglect, based upon findings that RHD subjects with neglect typically do not perform as well as subjects without neglect on highly demanding tasks (Tompkins *et al.* 1994). The lack of meaningful relationship is most probably due to the extremely restricted range of neglect severity in the subject sample (BIT score range 133–145, with one extreme outlier score of 69).

an estimate of the reliability of the entire test (Anastasi 1988). Results indicated strong internal reliability for both NBD ($r_{nn} = 0{\cdot}85$) and RHD ($r_{nn} = 0{\cdot}96$) subjects.

Validity

Spearman correlations were calculated to estimate convergent validity (WM with backward digit recall) and divergent validity (WM with forward digit recall, WM with word recall). Results are shown in table 3. For both administrations of the WM task, convergent validity correlations were fairly high for the NBD group but quite low for the RHD subjects. Divergent validity correlations were moderate to strong for both groups. However, there were large differences in effect sizes (from 12% to 32%), calculated by squaring the correlation coefficients, for the two presentations of the WM measure.

Discussion

Reliability

Test–retest reliability

The test–retest results indicate that the current measure is fairly stable over time for both subject groups. The stability of this task for RHD subjects is encouraging. Combined with predictive validity data indicating relationships between WM and performance on computationally demanding comprehension tasks (Tompkins *et al.* 1994), these results suggest that the WM measure may be useful as a prognostic indicator or an index of severity for this clinical group.

There are no test–retest reliability data reported for WM tasks in the normal ageing literature, but the test–retest reliability of the current measure is higher than that reported in similar studies with young adults. In one such study, Waters and Caplan (1996) evaluated reliability over time for their own WM reading span and Daneman and Carpenter's original reading span tasks. The amount of time elapsed between time 1 and time 2 was not carefully controlled, with a range of 31–176 days ($M = 84$ days). The results indicate correlations of $r = 0{\cdot}65$–$0{\cdot}66$ for two components of the Waters *et al.* (1987) task and a smaller relationship, $r = 0{\cdot}41$, for Daneman and Carpenter's reading span. However, these correlations may be confounded in unknown ways by the large and variable amount of time between the two testing sessions.

Although test–retest reliability was respectable for the current measure, error variability prevented a perfect correlation between time 1 and time 2. Possible sources of this error variance can be evaluated at the individual and the group level. The effect of practice is one potential factor in performance differences over time. The lack of group differences in WM performance over time suggests that practice did not play a major role in this study. At the individual level, a practice effect was operationally defined as improvement greater than 5%, or more than three points. Based upon this criterion, only four NBD and one RHD subject evidenced an effect of practice. The NBD subjects consisted of three females and one male who were no different from the larger group in terms of age, years of education, or digit and word recall scores. The RHD subject who demonstrated a practice effect was similar to the RHD group in terms of age, education, and backward digit span. His

scores for forward digit recall and word recall were slightly higher than the group average, but he had the greatest amount of time elapse between time 1 and time 2 (26 days). Based upon these data, there do not appear to be any specific characteristics which differentiate the individual subjects who demonstrated a practice effect from the larger groups.[4]

Another factor which could influence error variance in performance over time involves differences in administration or procedures across testing sessions. The majority of NBD subjects were tested by two different experimenters (because these subjects were participants in two different studies, as described in the Methods section). However, both experimenters were highly practiced in administering the WM task and the stimuli were presented from an audiotape recorder; thus, any such differences are likely to be minimal.

Internal consistency

Internal consistency of the current measure was impressive. The results from the current WM task are nearly identical to those reported by Salthouse and Babcock (1991) for their listening span task administered to a large group of adults (N = 227) of all ages (20–87 years old). No other study of older adults provides information regarding the internal consistency of WM tasks used. The high correlation between the two halves of the test indicate that they measure a common entity (Anastasi 1988, Norusis 1990).

Validity: NBD subjects

Validity evidence for NBD subjects will be discussed in relation to results reported in the few available studies that have examined the validity of various WM measures. For purposes of comparison, correlational data from this study and those from other investigations are summarized in table 4.

Convergent validity

A strong correlation between WM and backward digit recall performance was expected, due to the simultaneous demands of processing and retention required by each of these tasks. The moderate correlations obtained in the current study suggest that there is some overlap in the abilities measured, but the processes involved are not identical, and the tasks are not redundant (Anastasi 1988). Because the backward digit task traditionally has been considered to assess WM, the relationship between it and the WM task is consistent with the hypothesis that the current task assesses WM abilities. Although the storage components for the WM and backward digit recall tasks are probably similar, the processing required for language comprehension may be different from that of reversing serial order of digits. The relationship between the two tasks, then, may be due to similarity of overall demands, not necessarily the similarity in processing requirements.

[4] Anecdotally, at the end of the final session, subjects were asked if they remembered completing the WM task before, or if they recalled any specific items from the test. While the majority of the subjects reported doing a similar task, most of them did not think it was the same task both times, and only five subjects recalled specific items from the first presentation. None of the subjects who demonstrated a practice effect recalled any specific items from the task.

Table 4. Summary of correlations between WM[a] and other recall tasks for non-brain-damaged older adults

Study	Backward digit span[b]	Computation span	Forward digit span[b]	Word span[b]
Current study				
Time 1	0·69	–	0·52	0·59
Time 2	0·68	–	0·77	0·68
Light and Anderson (1985)				
Group 1	0·27	–	0·24	0·17
Group 2	0·33	–	0·40	0·44
Salthouse and Babcock (1991)[c]				
Group 1	–	0·68	0·45	0·56
Group 2	–	0·49	–	0·62
Wingfield *et al.* (1988)	–	–	0·30	0·40

[a] All studies report WM span scores, except for the current study, which used error scores.
[b] The current study reported error scores, not span scores.
[c] Subject groups comprised a mixed sample of young and older adults.

There have been very few studies in the ageing literature which have compared performance on two tasks purporting to measure working memory ability and the results from these studies are inconsistent (table 4). The correlations provided in the table should be interpreted with caution, because convergent validity was examined using two non-standardized measures. The current results are similar to those of Salthouse and Babcock (1991), who reported a moderately strong correlation ($r = 0{\cdot}68$) between performance on an auditory listening span task and a computational span task for their mixed group of young and older adults. The computational span task required subjects to select answers to simple arithmetic problems (arranged into sets of 1–7 problems) and recall the final digit of each problem after each set was completed. The WM and computational tasks were designed to be structurally similar to allow direct comparisons, with the only major differences between the tasks being the type of processing required (linguistic vs. mathematical).

In contrast, Light and Anderson (1985) computed correlations between working memory span and backward digit recall, but found no meaningful relationship between the two tasks for older adults. The characteristics of the working memory task employed by these authors may have exaggerated the discrepancies between the measures. Their working memory task was a reading span, whereas the digit span task was presented auditorily. Perhaps more importantly, the length of their reading stimuli was much greater (range 6–12 words) and less consistent across items than those used in the current WM task. Additionally, the authors calculated only span scores, which may have diminished the correlation coefficients due to the restricted range of scores.

Divergent validity

A moderate relationship was obtained for forward digit recall and WM scores in this study. This relationship was stronger than anticipated, a finding that could be related to the simplicity of the current WM stimuli. Other WM task variations used with older adults consist of sentence stimuli of at least 6–16 words. These auditory

span tasks also typically use sentences that require academic knowledge (e.g. statements taken from history/geography quiz books). For example, Wingfield *et al.* (1988) used statements from a quiz book, which were presumed to be of moderate difficulty. These items appeared to be heavily dependent upon academic knowledge (e.g. 'Dean Rusk was the Soviet Premier during the Cuban missile crisis'). The current study's use of short, simple sentences of general world knowledge may have lessened the amount of linguistic processing and academic knowledge necessary, thus making the current task more similar to simple STM tasks.[5] This speculation is consistent with the higher mean WM span scores for the current task ($M = 5{\cdot}3$) than those reported elsewhere in the ageing literature (M approximately 3·0; Light and Anderson 1985, Salthouse and Babcock 1991).

Another potential problem with generalizability is demonstrated in all of the studies that have gathered validity data across time or subject groups. Table 4 indicates that Light and Anderson (1985) obtained a large difference in the strength of correlations between WM and forward digit span for two groups tested on the same tasks [experiment 1: $r = 0{\cdot}24$ ($N = 25$), experiment 2: $r = 0{\cdot}40$ ($N = 20$)]. Our results, using two administrations of the WM task with the same subjects, also are quite divergent. This raises the issues of replicability and how confidently we can interpret results from giving any test or measuring any behaviour just once. Perhaps the common practice of accepting results from only one test administration of any non-standardized measure is not the most appropriate. This problem is not a new one, nor is it unique to our discipline. The results of this study merely remind us again to be aware of it. Of course, poor reliability of either the digit span or WM task also could contribute to inconsistent results across different subject groups.

The moderately strong relationship between WM and simple word recall for NBD subjects in this study fall at the high end of correlations reported in the ageing literature (table 4). One divergent result ($r = 0{\cdot}17$) was reported by Light and Anderson (1985). However, this low correlation may have been due to the questionable replicability across subjects (as noted above); particularly because the authors reported a moderate relationship ($r = 0{\cdot}44$) for a second group of subjects tested on the same tasks.

The moderate relationship between word recall and WM for NBD subjects in this study is slightly higher than the expected outcome for this study. This may reflect both the memory component shared between the two tasks and the fact that our WM task was simpler than those used in other investigations. However, because the correlations are only moderate in strength, there appears to be another component which differentiates the tasks. This is postulated to be the processing component of the WM task, which is not required for simple word recall (particularly when correct serial order is not required for the word recall task). This result is consistent with the proposal that the WM listening span paradigm taps more than simple STM; however, as discussed above, the simplicity of the current WM task may make it a less sensitive measure of concurrent processing and storage.

[5] Despite the relative simplicity of the current WM task as compared to other WM measures, the task clearly was not easy for the subjects. No subject obtained a perfect score across task administrations. Subject reports and experimenter observations also indicated that subjects had to focus intently on the task in order to complete it successfully.

Validity: RHD subjects

Convergent validity

In contrast to the NBD group, there was no meaningful relationship between WM and backward digit recall for RHD subjects. One possible reason for this result is the relatively low internal consistency of the backward digit recall task as administered in this study. The weaker reliability of this task will reduce the strength of the expected correlations (Anastasi 1988). Additionally, the small number of subjects in this group restricts the confidence in the results obtained.

Divergent validity

The relationship between WM and forward digit recall error scores for RHD subjects was similar to that reported for NBD subjects, probably for similar reasons. It is unclear why the RHD group demonstrated such a strong relationship between word recall and WM. The results diverge from those reported by Tompkins and colleagues (1994), who found no meaningful relationship between WM errors and simple word span ($r < -0{\cdot}30$). Examination of the subject groups tested in these two studies failed to reveal any obvious differences in demographic characteristics or WM ability that could have influenced the results. Differences in the testing conditions may account for some of the discrepancy in results. In the Tompkins *et al.* study, the word span task was administered after all other testing was completed (the number of weeks elapsed between the WM and word recall task administrations was unspecified) and was presented live-voice over the telephone. Additionally, the use of span scores, as mentioned previously, may have lowered the correlation coefficient (Anastasi 1988).

The strong correlations between working memory and the simple recall tasks for the RHD group may reflect a strategic response to the dual demands of the working memory task. When these subjects had difficulty carrying out simultaneous processing and storage requirements, they may have put more effort toward the true/false component, which demanded an immediate response, at the expense of storing or rehearsing sentence-final words for recall. If so, they would treat the working memory task as a single operations task, making it more similar to the short-term memory measures in terms of processing demand. This effect would not occur with the NBD subjects, because they were able to complete the task more successfully; their WM capacities were not taxed to the point which might induce this effect. One way to investigate whether or not subjects were using a strategy such as this would be to reverse the task requirements, having subjects repeat each sentence-final target word as the stimuli are presented, then recalling in order the true/false value of each sentence in a single set.

The relatively small samples tested in this study may be another factor influencing the strength of relationships for all tasks. As was seen with the results reported by Light and Anderson (1985), there is a danger in generalizing from one small sample to another—even when the tasks used are reliable. Differences in subject characteristics influence the results and without a large, representative sample, any generalization must be made with caution.

Conclusions

To summarize our findings, the current working memory task is internally consistent and reliable over time. It is also more reliable than the other tasks currently used in the literature. The validity data are less clear and raise questions about what this version of the working memory task actually measures. The validity results for the NBD group indicate that the task is similar to another purported working memory task and that it can be contrasted with measures of simple short-term memory. However, the differences between the convergent and divergent correlations are not as large as anticipated. Validity data for the RHD subjects are even more difficult to interpret. The small sample size and low internal consistency of the backward digit recall task may play a role in the weak relationship between this task and the working memory measure.

Although non-standardized measures may appear to be the easiest, or most appropriate tools, interpretation of results is suspect, cross-study comparisons are difficult, and generalization can be only tentative, at best. Despite the problems evident in the use of non-standardized tools, it is obviously not reasonable to insist that all clinical and research tools must be standardized. However, appropriate interpretation and generalization of results can be facilitated by estimating the reliability and validity of such tools (Tompkins 1992, Willmes 1990). We also may wish to express explicitly the cautions connected with using them.

The present study is the first step toward improving one commonly used task paradigm. Despite the questions that we are still wrestling with, our results suggest that this step is in the right direction. We will continue to test subjects on these tasks and will collect data on a more heterogeneous group of RHD subjects. Additionally, we will be gathering more predictive validity data. With data from a larger, more heterogeneous, sample of subjects and more information about how performance on this task is related to language comprehension, we hope to be able to establish a clearer picture of the construct measured by the current working memory measure.

References

ANASTASI, A. 1988, *Psychological Testing* (6th edn) (New York: Macmillan).

ANNET, M. A. 1970, A classification of hand preference by association analysis. *British Journal of Psychology*, **61**, 303–321.

BAYLES, K. A. and TOMOEDA, C. K. 1990, *Arizona Battery for Communication Disorders of Dementia* (Tucson, AZ: Canyonlands Publishing).

BENTON A., HAMSHER, K. DE S., VARNEY, N. and SPREEN, O. 1983, Judgment of line orientation. *Contributions to Neuropsychological Assessment* (New York: Oxford University Press), pp. 44–54.

BROOKSHIRE, R. H. and NICHOLAS, L. E. 1993, *Discourse Comprehension Test* (Tucson AZ: Communication Skill Builders).

COHEN, J. and COHEN, P. 1983, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd edn) (Hillsdale, NJ: Erlbaum).

DANEMAN, M. and BLENNERHASSETT, A. 1984, How to assess the listening comprehension skills of prereaders. *Journal of Educational Psychology*, **76**, 1372–1381.

DANEMAN, M. and CARPENTER, P. A. 1980, Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, **19**, 450–466.

DANEMAN, M. and CARPENTER, P. A. 1983, Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9**, 561–584.

DUNN, L. M. and DUNN, L. M. 1981, *Peabody Picture Vocabulary Test—Revised* (Circle Pines MN: American Guidance Service).

FOLSTEIN, M. F., FOLSTEIN, S. E. and MCHUGH, P. R. 1975, Mini-mental state: a practical guide for grading the cognitive status of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189–198.

GOODGLASS, H. and KAPLAN, E. 1983, *Assessment of Aphasia and Related Disorders* (2nd edn) (Philadelphia: Lea & Febiger).

JUST, M. A. and CARPENTER, P. A. 1992, A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, **99**, 122–149.

KUCERA, H. and FRANCIS, W. N. 1967, *Computational Analysis of Present Day American English* (Providence RI: Brown University Press).

LEZAK, M. D. 1995, *Neurological Assessment* (3rd edn) (New York: Oxford University Press).

LIGHT, L. L. and ANDERSON, P. A. 1985, Working-memory capacity, age, and memory for discourse. *Journal of Gerontology*, **10**, 737–747.

NORUSIS, M. J. 1990, *SPSS Base System User's Guide* (Chicago: SPSS).

SALTHOUSE, T. A. 1990, Working memory as a processing resource in cognitive aging. *Developmental Review*, **10**, 101–124.

SALTHOUSE, T. A. and BABCOCK, R. L. 1991, Decomposing adult age differences in working memory. *Developmental Psychology*, **27**, 763–776.

SEASHORE, C., LEWIS, D. and SAETVEIT, J. 1960, *Seashore Measures of Musical Talent* (New York: Psychological Association).

TOMPKINS, C. A. 1990, Knowledge and strategies for processing lexical metaphor after right or left hemisphere brain damage. *Journal of Speech and Hearing Research*, **33**, 307–316.

TOMPKINS, C. A. 1991, Automatic and effortful processing of emotional intonation after right or left hemisphere brain damage. *Journal of Speech and Hearing Research*, **34**, 820–830.

TOMPKINS, C. A. 1992, Improving aphasia treatment research: some methodological considerations. *Aphasia Treatment: Current Approaches and Research Opportunities* (Washington DC: National Institute on Neurological Communication Disorders), pp. 37–46.

TOMPKINS, C. A., BLOISE, C. G. R., TIMKO, M. L. and BAUMGAERTNER, A. 1994, Working memory and inference revision in brain-damaged and normally aging adults. *Journal of Speech and Hearing Research*, **37**, 896–912.

TOMPKINS, C. A., BOADA, R. and MCGARRY, K. 1992, The access and processing of familiar idioms by brain-damaged and normally aging adults. *Journal of Speech and Hearing Research*, **35**, 626–637.

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

WATERS, G. S. and 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. and HILDEBRANDT, N. 1987, Working memory and written sentence comprehension. In M. Coltheart (Ed.) *Attention and Performance XII: The Psychology of Reading* (Hillsdale, NJ: Erlbaum), pp. 531–555.

WATERS, G. S., CAPLAN, D. and ROCHON, E. 1995, Processing capacity and sentence comprehension in patients with Alzheimer's disease. *Cognitive Neuropsychology*, **12**, 1–30.

WECHSLER, D. 1987, *Wechsler Memory Scale—Revised* (Manual) (San Antonio: The Psychological Corporation–Harcourt Brace Jovanovich).

WILLMES, K. 1990, Statistical methods for a single-case study approach to aphasia therapy research. *Aphasiology*, **4**, 415–436.

WILSON, B. A., COCKBURN, J. and HALLIGAN, P. 1987, *Behavioural Inattention Test* (Bury St. Edmunds, Suffolk, UK, Thames Valley Test Company).

WINGFIELD, A., STINE, E. A. L., LAHAR, C. J. and ABERDEEN, J. S. 1988, Does the capacity of working memory change with age? *Experimental Aging Research*, **14**, 103–107.

Appendix: Working memory stimuli

Level 2

Set 1	Set 2	Set 3
You sit on a chair. (T)	A table is an animal. (F)	Tigers live in houses. (F)
Trains can fly. (F)	Children like games. (T)	Milk is white. (T)

Level 3

Set 4	Set 5	Set 6
Sugar is sweet. (T)	You ride on a bus. (T)	Pumpkins are purple. (F)
Florida is next to Ohio. (F)	Cats can talk. (F)	Mice are smaller than lions. (T)
Horses run in the sky. (F)	Apples grow on trees. (T)	Roses have thorns. (T)

Level 4

Set 7	Set 8	Set 9
Twelve equals one dozen. (T)	Water is dry. (F)	Chickens eat eggs. (F)
Bicycles are slower than cars. (T)	Cows like to eat grass. (T)	Babies can drive. (F)
A book can play. (F)	Ducks have webbed feet. (T)	A clock tells time. (T)
Feathers can tickle. (T)	Little boys wear dresses. (F)	The sky is green. (F)

Level 5

Set 10	Set 11	Set 12
Carrots can dance. (F)	An orange is a fruit. (T)	You keep books in ovens. (F)
Fish swim in water. (T)	February has sixty days. (F)	Rabbits can read. (F)
You sleep on a bed. (T)	A shoe has ears. (F)	A lobster has a shell. (T)
You eat breakfast at night. (F)	You wash with soap. (T)	Chairs can eat. (F)
People have eyes. (T)	A car can race. (T)	Dogs have four legs. (T)

Recall targets are underscored. T = true, F = false.