Measuring episodic memory: A novel approach with an indefinite number of alternative forms

Michael Humphreys[1], Simon Smith,[1] Nancy A. Pachana[1], Gerry Tehan[2], Gerard J. Byrne[3]

[1] School of Psychology, University of Queensland, Australia
[2] School of Psychology, University of Southern Queensland, Australia
[3] School of Medicine, University of Queensland, Australia

Requests for Reprints:

Professor Michael Humphreys
School of Psychology
University of Queensland
Brisbane, Qld 4072
Australia
mh@humanfactors.uq.edu.au

Authors' final corrected pre-publication version of:
Humphreys, Michael S. and Smith, Simon and Pachana, Nancy A. and Tehan, Gerald and Byrne, Gerard J. A. (2010) *Measuring episodic memory: A novel approach with an indefinite number of alternative forms.* Applied Cognitive Psychology, 24 (8). pp. 1080-1094. ISSN 0888-4080. Accessed at USQ ePrints http://eprints.usq.edu.au/5593/

Abstract

Both clinical practice and clinical research settings can require successive administrations of a memory test, particularly when following the trajectory of suspected memory decline in older adults. However, relatively few verbal episodic memory tests have alternative forms. We set out to create a broad based memory test to allow for the use of an essentially unlimited number of alternative forms. Four tasks for inclusion in such a test were developed. These tasks varied the requirement for recall as opposed to recognition, the need to form an association between unrelated words, and the need to discriminate the most recent list from earlier lists, all of which proved useful. A total of 115 participants completed the battery of tests and were used to show that the test could differentiate between older and younger adults; a sub-sample of 73 participants completed alternative forms of the tests to determine test-retest reliability and the amount of learning to learn.

Measuring Episodic Memory with an Indefinite Number of Alternative Forms

Standardized tests of episodic memory are used for a wide variety of clinical and research purposes. With an ageing population increased attention is being directed to early detection of cognitive decline in older individuals. Differentiation between normal age-related changes in memory and declines suggestive of either underlying neuropsychiatric or pathological changes is vital. Monitoring behavioural or pharmacological interventions aimed at improving memory performance is also clinically important. A memory test which could be used for monitoring purposes across time would be of great utility in clinical settings, particularly primary care settings, as well as research contexts including longitudinal studies. Neuroimaging studies also require memory tests which ideally have multiple alternate forms and are easy to administer while participants are being scanned. Finally, the ability to administer a memory test in a standardised and replicable manner has implications for outreach to remote patients via e-health and telemedicine approaches.

These examples ideally require the repeated administration of a memory test. However, this raises the issue of practice effects. While an increasing number of neuropsychological tests offer limited alternate form options, this is far from the norm, and many tests offer only a single alternate form which does not address the issue of repeated testings required in some clinical and research settings (Brandt, 1991; McCaffrey & Westervelt, 1995). For example, Automated Neuropsychological Assessment Metrics (ANAM) measures have utilised computers to provide alternate forms (Kane, Roebuck-Spencer, Short, Kabat & Wilken, 2007). CogState, a popular automated test used in drug trials, has demonstrated sensitivity to early cognitive impairment in several studies (e.g. Maruff et al., 2004). Yet such batteries' grounding in cognitive psychology theory arguably is limited.

From a cognitive psychology perspective there are three main difficulties involved in creating and using an episodic memory test with multiple alternate forms in order to track performance over a period of time. The first is how to measure episodic memory. Will any episodic memory task do or are some tasks better than others? Will a single task suffice or is it necessary to measure episodic memory using multiple tasks?

Item differences pose the second problem. Although memory researchers do not routinely test for both subject and item differences, the investigation of item differences has at times played an important role in memory research (Underwood, 1975). One can subdivide item differences into a main effect due to items and into a subject by item interaction. One can eliminate the main effect by testing different sets in order to ensure their comparability. This, however, is an expensive process and will not remove subject by item interactions. These interactions do not have to be removed if the focus is on group differences, especially if individuals have been randomly assigned to groups. However, they may be a problem if naturally occurring groups are used and they are definitely a problem if the focus is on monitoring changes within an individual.

The final problem in repeatedly using a memory test to monitor changes in performance is learning-to-learn. Performance on many memory tasks improves with practice even when different items are being learned (Postman & Schwartz, 1964). Any such performance changes will have to be taken into consideration in interpreting a pattern of changes with repeated administrations of a memory test.

In the current study a broad test of episodic memory with multiple components was created. The use of multiple tasks provides some insurance against the possibility that different neurological substrates are involved in at least some tasks commonly considered episodic (Norman & O'Reilly, 2003). The tasks created require recall as well as recognition, with some requiring formation of associations between pairs of unrelated words (see Lowndes & Savage, 2007; for a justification for using tasks which have this requirement in the early detection of memory impairment in Alzheimer's disease) , while others involve discriminating words on recent lists from earlier lists. These three variables were combined to create four tasks. Two tasks required recall and two required recognition. One of the recall tasks and one of the recognition tasks required subjects to form associations between arbitrary pairs of words. In addition, 3 out of the 4 tasks had separate scores for 2 or more task subcomponents. The initial objective was to evaluate each of the tasks and subcomponents with respect to psychometric properties and clinical utility. Different components might also serve different functions, for example some might primarily reflect memory performance whereas others might primarily reflect attention and the following of instructions.

Experimental Plan

There were four goals for this first stage of test development. First we wanted to ensure that our test provided broad coverage of the concept of episodic memory. We thus wanted the inter-correlations between the four tasks to be positive but of only moderate strength. If the correlations between two of the tasks were too high it would be an indicator that the two tasks were measuring essentially the same construct and if they were not positive it would be an indicator that they were measuring different constructs. Two additional memory tasks were also administered to a sub-sample of our participants in order to evaluate the breadth of the test.

Second, in order to provide what was an essentially unlimited number of alternative forms, individual tests were constructed by randomly sampling words from a larger pool of words. The words were mostly two syllables and of intermediate frequency. Words which could be auditorially confused were eliminated. We thus needed to know if we could achieve satisfactory levels of test-retest reliability with this procedure.

Third, because the test was designed so that it could be repeatedly administered we were concerned with the amount of learning to learn which would occur. If participants performed better on the second and subsequent test administrations it would be necessary to take the expected improvement into consideration in constructing the test norms.

Fourth, in a preliminary attempt to validate the test for the study of individual differences, we sought to show that the test was negatively correlated with chronological age. Chronological age was chosen because the participants were readily available and there is good data on the effect size of different tasks (La Voie & Light, 1992).

Method

Participants

The 115 participants comprised 45 younger (< 40 years) and 70 older (>60 years) adults residing in Brisbane, Australia. The younger adults were university students and staff recruited through advertisement and electronic newsletter. Older adults were community dwelling, and were recruited through advertisements in local environs, electronic newsletter and through community newspapers. The first 36 participants recruited came from the University and were tested in a single session.

After their results had been analysed it was decided to make some changes to Task 1. The remaining 79 participants were then recruited. These came from both the university and the wider community. Of these 6 were tested on one occasion and the remaining 73 were tested on three separate occasions each. The 73 participants who were tested on three occasions were used to assess test retest reliability and learning to learn. They also provided the comparisons with the other memory tasks. The total group of 115 was used to assess the relationship with aging. Participants were informed that they would be completing some short memory tasks on a computer, and would be completing demographic and mood questionnaires. Participants with no problems viewing and using a computer were included, and the use of glasses or a hearing aid was accepted. Participants were compensated for their time with a coffee voucher and a snack.

Procedure and Materials

All participants attended the University of Queensland School of Psychology for individual testing sessions. In the first part of the session, participants filled out a questionnaire booklet designed to assess general demographic and health characteristics. In the second part, the experimenter told the participants that they would see instructions on the computer screen, and that a series of three short tasks would follow. Participants then completed the tasks in a sound attenuated cubicle. The tasks were presented on an IBM-compatible PC. Each task was preceded by on-screen text instructions. Tasks were presented in a fixed order (1 to 4) for each participant. Participants began each task in their own time by initiating the task with a button press, and could take short breaks between tasks if necessary. Participants responded verbally and these verbal responses were recorded on sound files for subsequent scoring. Verbal responding was used because we thought that it would be easier for our elderly sample. Scoring from recordings is also more accurate than on line scoring. The testing sessions lasted 45 to 60 minutes and research staff were available to answer any questions or concerns that participants had about the tasks.

Of the 115 participants, 73 participants completed the 4 tasks in the battery on an initial occasion, then completed alternative forms of the battery on 2 further occasions, at one week intervals after the initial test session. On the second testing they also completed an additional brief paired associates task administered by the researcher, and on the third testing they completed a computer administered operation

span task. The remaining 42 participants completed the 4 tasks in the battery on a single occasion.

Measures:

Verbal paired associates task. Participants in the test-retest arm of the study were administered the Verbal Paired Associates (VPA) subtest of the Wechsler Memory Scale – Third edition (WMS-III, The Psychological Corporation, 1997), with no delay trial. This subtest is one of the most widely used instruments for assessing explicit episodic memory performance (Uttl, 2005), and.requires participants to learn eight unrelated word pairs across four study-test trials

Operation Span Task. In this task, participants were shown words on the computer screen, one at a time. A simple maths problem was presented beside each word (e.g. 4x2+7=15). Partipants were instructed to read each word out aloud when they saw it, and also to read the numbers in the math problem out aloud. The word and equation were presented for 6 seconds. Participants were then instructed to repeat the words that they had read (not the numbers) at the presentation of an on-screen cue. After four practice trials, participants were presented with four trials of a two word list, then four trials of a three word list, and finally four trials of a four word list. The score is the total number correct across all list lengths. (Turner & Engle, 1989; Tehan, Hendry & Kocinski, 2001).

Task 1. List Discrimination. The first group of 36 young participants studied two 60 word lists. Forty of the words in each list were presented once and 20 were presented twice so there were 80 presentations in each list. Words were presented at a two sec rate; List 2 immediately followed list 1 and the test immediately followed List 2. On the test all 120 words (60 from each list) were displayed and people were asked to indicate whether the word occurred in List 2. Performance of the initial participants was quite poor so the task was changed for the remaining participants. In the new version the lists contained 40 words, 20 presented once and 20 twice. The retention interval between List 1 and List 2 was standardized at 30 seconds. The test contained 80 words (40 from each list).

The twice presented words were included so that participants could not respond simply on the basis of familiarity. Note that regardless of how familiarity is conceived the task is designed to require the use of list specific information just as in the exclusion condition of the process dissociation procedure (Jacoby, 1991). In fact

it seems quite likely that performance on this task will correlate quite highly with estimates of recollection obtained from the process dissociation procedure.

Experimental Task 2: Pair Recognition. In Task 2 participants studied 40 pairs of words presented at a 4 second rate. They were instructed to learn the pairs so that they could recall one member if they were given the other member as a cue. The test started immediately after the end of the study list. People were shown 20 intact pairs (the pair was one they had studied) and 20 rearranged pairs (two old words but not studied in the same pair). They were instructed to indicate whether or not they had studied each test pair.

Experimental Task 3: Cued Recall. Participants were asked to study 20 4-pair lists. Each pair was presented for 4 seconds and participants were instructed to learn the pairs so that they could recall one member of a pair if they were shown the other member. The test started immediately after the presentation of the last study pair and consisted of the presentation of one member from each pair. The first cue came from the fourth (last) pair, the second cue came from the third pair, the third cue came from the second pair, and the last cue came from the first pair.

There have been several attempts to look at the effect of retention interval within conventional paired associate paradigms (Greeno, 1964; Izawa, 1971, 1972; Murdock, 1963) . At that time the thinking had been that retention at short intervals was very good because recall was coming from short term memory. More recently there have been doubts that there is a clean separation of short and long term memory (Mogle, Lovett, Stawski, & Sliwinski, 2008; Nairne, 2001; Tehan & Humphreys, 1995, 1996). In addition there has been an increasing interest in the interference generated by prior list items (Keppel & Underwood, 1962; Postman & Keppel, 1977; and from other items in the study list (Dennis & Humphreys, 2001; Ratclife, Clark, & Shiffrin 1990) and of the role of rapidly changing temporal cues (Glenberg, 1980; Howard & Kahana, 2002).

It thus seems likely that in the conventional paired associate paradigm the retention interval controls the amount of interference from other list items and from prior learning. That is, when a cued recall test immediately follows the study of a cue target pair the short retention interval will reduce the amount of noise from prior presentations of the other pairs from the study list and there are no presentations of other cue-target pairs intervening between study and test. When the retention interval is increased there will be an increase in the amount of noise from both prior and

subsequent cue target pairs. If this analysis is correct then our cued recall task is testing the same ability to form associations as is tested in a conventional multi-trial paired associate task.

Our task offers several pragmatic advantages over conventional paired-associate tasks. In the conventional procedure performance is frequently poor on the first trial or even the first few trials. This means that one must plan for multiple study and test trials which increases the time taken to administer the task and reduces the number of different words or word pairs that can be used. The small number of words and word pairs in turn increases the likelihood of item effects or subject by item interactions. In our task the number of study test pairs employed will reduce item effects and subject by item interactions, making it easier to construct alternative forms. Near ceiling performance on the fourth pair would also be an indication that subjects were following instructions and attending to the pairs.

Experimental Task 4: Immediate and Delayed Serial Recall. Task 4 was a variant on a task developed by Tehan and Humphreys (1995, 1996). Test takers were presented with one or two blocks of 4 words each. Block 1 was either followed by an immediate test of serial recall or the test taker was informed that they could forget block 1. The forget instruction was immediately followed by the presentation of the four words in block 2. Four seconds of distractor activity followed the presentation of block 2. This consisted of the presentation of eight digits which the test taker was asked to read out loud. After four seconds of reading digits the person was asked to recall the words from block 2 in serial order. All words were presented at a two second rate and on two-block trials there was a two second pause between block 1 and block 2 in which the forget instruction was presented.

The block 1 trials were designed to be relatively easy and to provide the test taker with a sense of accomplishment. The four seconds of interference activity at the end of block 2 was designed to prevent rehearsal and to provide a long enough retention interval, relative to the two second inter-block interval, so that there would be a substantial amount of interference from the first block.

Results

The results from the first testing session for participants who had repeated tests and from the only testing session for the other participants are given in Table 2. By

inspection it can be seen that performance on the fourth pair of Task 3 (the first pair tested) is very good for all groups of participants. This is an indication that participants across all groups understood the instructions and were paying attention. A similar result occurred with the immediate test of Task 4. Again the data suggest participants in all groups understood the instructions and paid attention. These two results are important indicators that participants are performing in the expected manner. However, there is a ceiling effect present with both scores which severely restricts the range. The only other notable point is that Task 1 was quite difficult, especially for the older group, and there is a consequent floor effect.

Insert Table 2 about here

Correlation between Tasks & Correlation with age:

To provide an estimate of the convergent validity of the task battery, the correlations between the four tasks were calculated using the data from the first testing session for all 115 participants. To examine the influence of age on these measures, correlations between each task and age were also calculated (Table 3). Individual scores on each task were transformed to z-scores (individual score-total sample mean/total sample SD).

Insert Table 3 about here

Almost all of the inter-correlations among the tasks and the task components were significant though moderate in size. The exceptions were the two conditions which have been identified as displaying ceiling effects (Pair 4 Task 3 and Immediate Test Task 4). In addition d' for non repeated items in Task 1 failed to correlate significantly with the Delayed Test on Task 4. The moderate correlations between the different episodic memory tasks support the idea that there is a lot of specific task

variance in the measurement of episodic memory. Further support for this idea comes from the correlation of .76 between the Pair 2 and 1 scores in Task 3, which is the single highest correlation in the Table. In addition, the correlation between the Immediate and Delayed test in Task 4 of .52 was the single largest correlation involving the immediate test. This is especially impressive given the restriction on the range of the Immediate Test scores.

Significant correlations with age were found for all task components, with the exception of the two tasks where a ceiling effect had been identified (Pair 4 Task 3 and Immediate Test Task 4). The largest correlations with age were with Task 3 (Pairs 2 and 1). All of the other correlations with age were more moderate.

Effect Size estimates:

The ability of the tasks to discriminate between younger and older adults was further examined by estimating effect sizes for age differences in task performance (Table 4). The comparison groups for this analysis comprised a group of 45 young adults (18-40 years old, mean age 20.3, SD 5.0) and a group of 46 healthy, community dwelling older adults (60+ years old, mean age 67.9, SD 6.9). Participants aged 40-60 were excluded from this analysis. Standardized effect size estimates for each task metric (g) were corrected to provide unbiased estimates of effect size (d) following the procedures detailed in Light et al. (2000) and Hedges and Olkin (1985). The 95% confidence intervals for d suggest significant age differences on all tasks with the exception of Pair 4 Task 3. Mean significant effect size estimates ranged from 0.49 to 1.32 (Table 3). Note that 36 of the younger adult sample completed an alternative version of Task 1 (increased list length and unstandardized inter-list interval), and age differences in this task should be interpreted with caution.

Insert Table 4 about here

Test-retest Analyses:

Learning to Learn. A sub-sample of 73 participants completed the task battery on three test occasions, one week apart (Session 1, Session 2 and Session 3). To test the effect of repeat administration of the measures on performance on each task, a series of repeated measures analyses were conducted. Test session (Session 1, Session 2 and Session 3) comprised three levels of the repeated factor. The mean and standard deviation for the task raw scores are presented in Table 5, with statistical significance and estimates of effect size noted. In addition, we have presented the results from the VPA which was collected in Session 2 and the OS which was collected in Session 3. Mauchly's test of sphericity was significant for Task 2 ($w=0.91$, $p=0.04$), Task 3, Position 3 ($w=0.81$, $p=0.00$), and for Task 4, immediate ($w=0.87$, $p=0.01$). Greenhouse-Geiser epsilon was used to correct the degrees of freedom for the standard ANOVA test in these cases, with minimal effect on the reported significance. A significant increase in task scores was found for Pair 1 Task 3, Pair 2 Task 3, and for the Delayed Test Task 4. Eta-squared estimates of effect size suggested that 6-8% of the variability in these tasks scores was accounted for by the test session factor.

Insert Table 5 about here

It is impossible to tell from these results whether the increases in performance are due to inadequate understanding of the task on the part of a few participants in the first session or represent an increase in the ability to perform the task (e.g., in Task 3 this could represent a better way of forming an association between two unrelated words). Nevertheless the modest sizes of the changes indicate that this is not going to be a serious issue as this test is designed to track changes in memory functioning over multiple testing sessions. Task 1, however, may be an exception to this conclusion. In this task performance improved on the non-repeated items in the study lists and deteriorated for the repeated items. In order to test this we ran a supplementary analysis using a 2 (repeated vs. non-repeated items) by 3 (session 1 vs. session 2 vs. session 3) ANOVA. There was no main effects for trial type ($p=0.31$) or for session ($p=0.81$). The interaction approached significance ($p=.06$). It thus appears that for at

least some participants the strategy for performing the task may be changing over sessions.

Test-Retest Reliability. In order to examine test-retest reliability we calculated correlations between the scores for the individual tasks and their sub-components across the three test sessions. These are presented in Table 6. With respect to the test-retest reliability Task 1 is clearly an outlier. One reason for the low reliability of this task is the high level of difficulty with a consequent floor effect. In addition, this was the task where the relative difficulty of the two components may have changed over the three test sessions. It thus seems highly likely that strategies for performing this task are changing over the sessions. Perhaps some participants started out with the idea that they could simply choose the most familiar item and then changed to a new strategy when they realized that this was not working. Pairs 4 and 3 in Task 3 also showed little in the way of test-retest reliability from session 1 to session 2. This is understandable in the case of Pair 4 where performance was very close to ceiling. It is less understandable with respect to Pair 3 where performance was not on ceiling.

The test-retest reliability of the Pair 3 scores was better between sessions 2 and 3. However, there was no converging evidence on the question of a strategy change as Pair 3 performance stayed nearly constant across the three testing sessions.

Insert Table 6 about here

Global Scores

Based on the results from the previous analyses it was clear that Task 1 would not make an appropriate contribution to an overall test score (low test retest reliability, evidence for strategy changes, floor effects). In addition, position 1 of Task 3 and the immediate condition of Task 4 had ceiling effects with position 1 of Task 3 also exhibiting low test retest reliability. We thus excluded these scores and used the remaining scores to calculate a composite (Global Score). This was calculated as the mean of individual z-scores on Task 2, Task 3 (mean of positions 2,3,4), and Task 4 Delay condition. Age differences in the Global score were estimated using the data and methods previously presented for effect size estimates. The standardized,

unbiased, estimate of effect size (d) was 1.04, with 95% confidence interval of 0.60-1.48.

The effect of repeat administration on the Global score was assessed in a repeated measures analysis of variance, with three levels of the test session factor. To provide an estimate of relative change, Global scores for each participant on each test session were calculated from z-score values for each of the sub-scales based on the distribution of sub-scale scores across the three test sessions. There was no increase in the mean relative Global score across the three test sessions ($F_{2,72}=0.029$, $p=0.97$, $\eta^2=0.00$).

In order to determine whether the Global Score measured a broad episodic construct we calculated the Global Score for each test session. We then calculated the correlations between these session specific Global Scores, the VPA scores from Session 2 and the OP scores from Session 3 (Table 7). The Global score correlated better with both the VPA and the Operations span test than the two did with each other. This pattern was quite stable across the three test sessions. This is another indication that we have created a relatively broad test of episodic memory.

Insert Table 7 about here

To assess the temporal reliability of the Global scores, test-retest correlations were performed on the Session 1, Session 2 and Session 3 data (n=73). Global scores were calculated from individual sub-scale z-scores based on the distribution of sub-scale scores on each test session (Table 8). Correlation coefficients (r) were significant in each case, with performance on one session accounting for 49-64% of the variance in the other two sessions. There are at least three sources of variability which could reduce the between session correlations. First, there is the instability inherent in a test which randomly samples items from a large pool. There is also the possibility that some participants will have misunderstood the test instructions and/or started the test with an inappropriate strategy. For example, in Task 3 (the paired associate task) it would be counterproductive to learn an association and/or to notice relationships between words in two different pairs. Finally, with the older participants, in particular, there might be some variation in their ability to cope with the demands of the testing sessions. For example, the time of day at which testing

occurred was not always standardized and quality of sleep the night before might have varied.

Insert Table 8 about here

We thus decided to see if the majority of participants were producing stable results with a minority showing a substantially greater amount of variation. We identified 8 outlier cases with standard deviation in Global score between the three test session >0.5. These cases were deleted from the analysis (n=65). Global scores were calculated from individual sub-scale z-scores based on the distribution of sub-scale scores on each test session (Table 9). Correlation coefficients (r) were significant in each case, with performance on one session accounting for 67-76% of the variance in the other two sessions. With the outliers removed, the correlation between session 1 and session 2 was now as large as the correlation between session 2 and session 3. This suggests that a small number of participants may have had difficulty with Session 1. This could have been a direct result of the negative affect (worry) induced by the unusual experience of participating in a psychology experiment in Session 1. It could also be an indirect effect where worry interfered with their comprehension of the instructions for one or more of the tasks.

Discussion and Conclusions

There were four goals for the first stage of test development. First, we wanted to create a test which provided broad coverage of the concept of episodic memory. The Global scale we constructed is that test and future applications will not use Task 1. We will continue to administer the test for position 1 in Task 3 and the immediate memory test in Task 4 even though the results on these subcomponents will not be included in calculating the Global score. We believe that performance on these tasks is primarily measuring attention and compliance with the instructions not memory. However, performance on these subcomponents will provide some indication as to whether a low score on the memory test is due to a memory problem or to an attentional/compliance problem. Second, because the test was designed for repeated administration we wanted to ensure that only a moderate amount of learning to learn would occur. Third, we needed to know if we could achieve satisfactory levels of test-

retest reliability when we formed tests using randomly drawn samples of words. Fourth, in a preliminary attempt to validate the test for the study of individual differences and within subject changes in performance, we sought to show that the test was negatively correlated with chronological age.

The correlations between the specific tests were positive though moderate in size. Most importantly the global score calculated from three of the tests correlated more highly with the Wechsler VPA and the Operations Span Test than these two tests did with each other. It thus appears that the test is reasonably broad. With respect to learning to learn, some of the individual components showed moderate amounts of learning to learn. However, the global score did not significantly improve over the three testing sessions. These sessions were separated by one week whereas in clinical practice the separation is likely to be considerably greater. It thus does not look like learning to learn will be a problem, even in research settings where one week between testing sessions is likely to occur.

Test retest reliability was problematic as the correlations were only moderate. We cannot be sure where the problem lies but there are a few hints in the data. First, for the global score the correlation between session 1 and Session 2 was somewhat lower than was the correlation between Session 2 and Session 3. This may indicate that there was some nervousness and/or failure to understand the instructions on the first session. This possibility receives some support from the increase in the test retest correlations when we discarded eight participants who had the largest between session variance. The reliability of Pair 3 (the second pair tested) was also lower than the reliabilities of Pairs 2 and 1. This finding is more compatible with a strategy change across sessions than with any fundamental unreliability due to the random selection of words to create the tests. McCaffrey and Westerveldt (1995) have recommended that when assessing the potential effect of an intervention, a test should be administered twice before the intervention commences so that the second administration can serve as the baseline measure. This recommendation was made in order to reduce practice effects (learning to learn). It is also likely to reduce strategy changes and increase the correlation between the baseline measure and the post intervention measure. Such a strategy would be easy to implement with the current test.

An additional constraint on test retest reliability may come from the composition of our sample. In our repetition sample 46 out of 73 participants were over the age of 60. The memory performance of an older sample such as this is known to be

susceptible to time of day effects (Hasher, Goldstein, & May, 2005). Memory performance in this group may also vary with the quality of the previous night's sleep and other temporary factors. This needs to be investigated but it is possible that an older sample will show more session to session variability on a memory test than a younger sample.

As expected, the test was negatively correlated with chronological age. In fact, the effect size for the global score (1.04) was very similar to the value estimated by La Voie and Light (1994) for recall measures (.97) in spite of the fact that it contained a recognition measure.

Conclusions and Future Directions

We feel confident that our experimental episodic memory protocol strikes a balance between the need for a reliable and repeatable test for use in clinical research, and the desirability of a test with a solid theoretical and methodological base. The proposed test is relatively brief, and appears well tolerated by a key population of interest, namely older adults. The test has been validated in clinical settings with individuals with MCI (Kingsbury, Pachana, Humphreys, Tehan, & Byrne, manuscript submitted). The test worked well in real-world clinical settings and was well-tolerated by patients.

Most importantly we have demonstrated that it is possible to create an overall memory test using multiple tasks and multiple lists within tasks. This allows us to achieve an adequate level of reliability even though we are randomly assigning items to tasks and lists. We feel such an advance has particular utility in tracking an individual's declines against his or her own baseline, and with reference to other intra-individual factors such as medications, health issues and mood states.

As the comparison between our younger and older participants shows, this level of reliability is clearly sufficient for group research. However, subsequent larger trials in clinical settings are required to ascertain the clinical utility of the test for identifying individuals whose memory performance was starting to deteriorate or for tracking changes within an individual (e.g., to monitor and document the effects of therapy). The use of the test for these purposes will have to proceed cautiously, and other research group's efforts are welcome. The test can be obtained in a beta form from the first author.

References

Brandt, J. (1991). The Hopkins verbal learning test: Development of a new memory test with six equivalent forms. *Clinical Neuropsychologist, 5,* 125–142.

Dennis, S. & Humphreys, M. S. (2001). A context noise model of episodic recognition memory. *Psychological Review, 108,* 452-478.

Glenberg, A. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning & Memory, 6*(4), 355-369.

Greeno, J. G. (1964). Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology, 67,* 286-295.

Hasher, L., Goldstein, D, & May, C. P. (2005). It's about time: circadian rhythms, memory, and aging. In Chizuko Izawa and Nobuo Ohta (Eds.) *Human learning and memory: Advances in theory and application: The 4th Tsukuba International Conference on Memory* (pp 199-217). Mahwah, NJ. Lawrence Erlbaum Associates.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269-299.

Izawa, C. (1971). Massed and spaced practice in paired-associate learning: List versus item distributions. *Journal of Experimental Psychology, 89,* 10-21.

Izawa, C. (1972). Retention interval hypothesis and evidence for its basic assumptions. *Journal of Experimental Psychology, 96,* 17-24.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language, 30,* 513-541.

Kane, R.L., Roebuck-Spencer, T., Short, P., Kabat, M., & Wilken, J. (2007). Identifying and monitoring cognitive deficits in clinical populations using Automated Neuropsychological Assessment Metrics (ANAM) tests. *Archives of Clinical Neuropsychology, 22, Suppl 1,* 115-126.

Keppel, G. & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning & Verbal Behavior, 1,* 153-161.

Kingsbury, R., Pachana, N.A., Humphreys, M., S., Tehan, G., & Byrne, G.J. Use of a computerized cognitive screen for memory in an MCI population. Under editorial review for *Australasian Journal of Rehabilitation Counselling.*

La Voie, D. &Light, L. L. (1994). Adult age differences in repetition priming: A meta-analysis. *Psychology & Aging, 9,* 539-553.

Lowndes, G. & Savage, G. (2007). Early detection of memory impairment in Alzheimer's disease: A neurocognitve perspective on assessment. Neuropsychological Review, 17, 193-202.

McCaffrey, R. J. & Westervelt, H. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review, 5,* 203-221.

Maruff, P., Collie, A., Darby, D., Weaver-Cargina, J., Masters, C., & Curriee, J. (2004). Subtle Memory Decline over 12 Months in Mild Cognitive Impairment. Dementia and Geriatric Cognitive Disorders,18, 342-348.

Mogle, J. A., Lovett, B. J. Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationships among working memory, secondary memory, and fluid intelligence. Psychological Science, 19, 1071-1077.

Murdock, B. B. (1963). Short term memory and paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 2,* 320-328.

Nairne, J. S. (2001). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53,* 53-81.

Norman K. A., O'Reilly R.C. (2003). Modelling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611-646.

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal Of Experimental Psychology: General, 106,* 376-403.

Postman, L. & Schwartz, M. (1964). Studies of learning to learn: I. Transfer as a function of method of practice and class of verbal materials. *Journal of Verbal Learning and Verbal Behavior, 3,* 37-49.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 1*6, 163-178.

Tehan, G. & Humphreys, M. S. (1995). Transient Phonemic Codes and Immunity to Proactive Interference. *Memory and Cognition, 23,* 181-191.

Tehan, G. & Humphreys, M. S. (1996). Cueing effects in short-term recall. *Memory & Cognition, 24*, 719-732.

Tehan, G, Hendry, L & Kocinski, D. (2001) Word length and phonological similarity effects in simple, complex and delayed serial recall task: Implications for working memory. *Memory, 9,* 333-348.

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

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30,* 128-134.

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment research and clinical practice. *Psychological Science, 16,* 460-467.

Table 1. Participant sub-sample demographic characteristics.

		Age			Gender		Education
Sample/sub-sample	N	Mean	SD	Range	Female	Male	<12 years
All	115	46.4	22.3	18-93	77(67%)	38(33%)	
Participants<40 years	45	20.3	5.0	18-40	33(73%)	12(28%)	0%
Participants 40-60 years	24	54.0	3.6	45-59	18(74%)	6(26%)	22%
Participants >60years	46	67.9	6.9	60-93	27(60%)	19(40%)	12%
Test-retest participants	73	60.6	9.5	26-93	46(63%)	27(37%)	10%

Notes: Demographic characteristics are provided for the entire sample and for the participant subgroups which contributed to the different analyses.

Table 2. Mean scores on tasks for analysis sub-samples and total sample.

		Younger (n=45)	Older (n=46)	Test-retest (n=73)	Total (n=115)
TASK	METRIC	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)
1	NonRepeat d'[1]	0.81 (0.61)	0.48 (0.51)	0.60 (0.61)	0.66 (0.63)
1	Repeat d' [1]	1.10 (0.75)	0.52 (0.50)	0.75 (0.71)	0.85 (0.74)
2	d'	1.45 (0.86)	0.93 (0.68)	1.08 (0.76)	1.18 (0.81)
3	Pair 4[2]	0.95 (0.08)	0.95 (0.09)	0.96 (0.07)	0.95 (0.09)
3	Pair 3[2]	0.71 (0.20)	0.58 (0.21)	0.61 (0.19)	0.64 (0.20)
3	Pair 2[2]	0.63 (0.26)	0.23 (0.20)	0.27 (0.21)	0.40 (0.29)
3	Pair 14[2]	0.62 (0.27)	0.28 (0.24)	0.27 (0.23)	0.40 (0.30)
3	Mean	0.66 (0.17)	0.51 (0.14)	0.53 (0.13)	0.57 (0.16)
4	Immediate[3]	0.91 (0.08)	0.84 (0.17)	0.86 (0.16)	0.87 (0.14)
4	Delayed[3]	0.72 (0.14)	0.58 (0.23)	0.58 (0.22)	0.63 (0.21)

[1]. Non-Repeat and Repeat d' are the d' scores for the once and twice presented study items in Task 1.
[2]. Pair 4 is the last study pair and the first test pair whereas pair 1 is the first study pair and the last test pair
[3]. The immediate and delayed recall conditions of Task 4

Table 3. Correlations between Task subscale scores (Z).

	Task 1		Task 2	Task 3				Task 4	Task 4
	non r	r	d'	Pair 4	Pair 3	Pair 2	Pair 1	immed	delay
Task 1 non repeat d'[1]	-								
Task 1 repeat d'[1]	**0.31**	-							
Task 2 d'	**0.23**	**0.26**	-						
Task 3 pair 4[2]	-0.10	0.01	0.07	-					
Task 3 pair 3[2]	0.06	**0.23**	**0.22**	**0.24**	-				
Task 3 pair 2[2]	0.15	**0.31**	**0.35**	-0.05	**0.39**	-			
Task 3 pair 1[2]	**0.21**	**0.33**	**0.38**	-0.05	**0.44**	**0.76**	-		
Task 4 immed[3]	0.05	0.13	0.17	0.15	**0.40**	**0.36**	**0.27**	-	
Task 4 delay[3]	**0.19**	**0.31**	**0.36**	0.04	**0.38**	**0.49**	**0.56**	**.52**	
age	**-.21**	**-.31**	**-.30**	.14	**-.28**	**-.66**	**-.57**	**-.17**	**-.31**

Notes: Bold <0.05, Grey=approaching 0.05

[1]. Non-Repeat and Repeat d' are the d' scores for the once and twice presented study items in Task 1.

[2]. Pair 4 is the last study pair and the first test pair whereas pair 1 is the first study pair and the last test pair

[3]. The immediate and delayed recall conditions of Task 4

Table 4. Effect size estimates for standardized age differences (Young Adults and Older Adults).

TASK	METRIC	d	95% Confidence Interval Lower limit	Upper limit
1	Non-R d'	0.56	0.14	0.98
1	Repeat d'	0.83	0.40	1.26
2	d'	0.64	0.22	1.06
3	Pair 4	-0.13	-0.54	0.28
3	Pair 3	0.62	0.20	1.04
3	Pair 2	1.32	0.86	1.77
3	Pair 1	1.12	0.68	1.56
3	Mean	0.87	0.44	1.30
4	Immediate	0.49	0.07	0.91
4	Delayed	0.66	0.24	1.08
	Mean of all effect sizes	0.70	0.27	1.12
	Mean of significant effect sizes	**0.79**	**0.36**	**1.22**

Table 5. Mean task scores on three test sessions.

		Session 1		Session 2		Session 3				
Task	Metric	Mean	SD	Mean	SD	Mean	SD	F	Sig.	η^2
1	Non-R d'	0.63	0.62	0.86	0.83	0.79	0.90	1.99	0.14	0.03
1	Repeat d'	0.77	0.71	0.62	0.70	0.72	0.77	0.98	0.38	0.01
2	d'	1.08	0.76	1.18	0.74	1.14	0.88	0.87	0.41[a]	0.01
3	Pair 4	0.96	0.07	0.98	0.05	0.99	0.03	**6.65**	**0.00**[b]	**0.08**
3	Pair 3	0.61	0.20	0.66	0.18	0.64	0.20	1.24	0.29	0.02
3	Pair 2	0.27	0.21	0.30	0.21	0.34	0.24	**4.32**	**0.02**	**0.06**
3	Pair 1	0.28	0.23	0.25	0.21	0.29	0.27	0.94	0.39	0.01
4	Immediate	0.87	0.16	0.89	0.14	0.87	0.18	1.58	0.21[c]	0.02
4	Delayed	0.59	0.22	0.63	0.20	0.63	0.21	**4.59**	**0.01**	**0.06**
VPA				15.81	8.30					
OS						30.48	4.11			

Notes: a: epsilon=0.92; b: epsilon=0.84; c: epsilon=0.87.

Table 6. Correlations between scores on each of three sessions for each task, and for the GDAS measures

		1-2	2-3	1-3
Task	Metric	r	r	r
1	Non-R d'	0.13	**0.27**	**0.26**
1	Repeat d'	**0.22**	0.14	0.15
2	d'	**0.67**	**0.66**	**0.59**
3	Posit 1	0.18	0.16	0.18
3	Posit 2	0.20	**0.45**	**0.31**
3	Posit 3	**0.49**	**0.54**	**0.52**
3	Posit 4	**0.48**	**0.52**	**0.65**
3	Mean posit 2,3,4	**0.53**	**0.62**	**0.68**
4	Immediate	**0.58**	**0.53**	**0.79**
4	Delayed	**0.67**	**0.82**	**0.77**
GDAS	Depression	**0.61**	**0.60**	**0.53**

Notes: 1-2, 2-3, and 1-3 = correlation between session 1 and session 2, between session 2 and 3, and between session 1 and 3, respectively. Bold indicates $p<.05$

Table 7. Correlation between Global Score (on each of 3 sessions), Verbal Paired Associates, Operation Span task, participant Age, Goldberg Depression and Goldberg Anxiety scores (on each of three sessions).

Session 1

data	Global	VPA	OS	AGE	Anxiety	Depression
Global	-					
VPA	**0.45**	-				
OS	**0.47**	**0.25**	-			
Age	**-0.26**	-0.14	**-0.25**	-		

n=72

Session 2

data	Global	VPA	OS	AGE	Anxiety	Depression
Global	-					
VPA	**0.45**	-				
OS	**0.46**	**0.25**	-			
Age	**-0.37**	-0.14	**-0.25**	-		

n=72

Session 3

data	Global	VPA	OS	AGE	Anxiety	Depression
Global	-					
VPA	**0.40**	-				
OS	**0.53**	**0.25**	-			
Age	**-0.23**	-0.14	**-0.25**	-		

n=72

Note that the Verbal Paired Associates (VPA) was only administered on Session 2 and the Operations Span (OS) was only administered on Session 3. The Global score the Anxiety score and the Depression score are from the indicated session. Bold indicates p<.05 (?)

Table 8. Test-retest correlations (r) for Global scores.

	Session 1	Session 2	Session 3
Session 1	-		
Session 2	**0.70**	-	
Session 3	**0.80**	**0.79**	-

All correlations significant at $p < .001$

Table 9. Test-retest correlations (r) for Global scores with outliers removed (N = 65).

.

	Session 1	Session 2	Session 3
Session 1	-		
Session 2	**0.87**	-	
Session 3	**0.82**	**0.85**	-

All correlations significant at $p < .001$