Archives of Clinical Neuropsychology Advance Access published December 2, 2011

OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Immediate List Recall as a Measure of Short-Term Episodic Memory: Insights from the Serial Position Effect and Item Response Theory

Brandon E. Gavett[1,2,*], Julie E. Horwitz[3,4]

[1]*Department of Psychology, University of Colorado at Colorado Springs, Colorado Springs, CO, USA*
[2]*Department of Neurology and Center for the Study of Traumatic Encephalopathy, Boston University School of Medicine, Boston, MA, USA*
[3]*Psychology Service, Edith Nourse Rogers Memorial Veterans Hospital, Bedford, MA, USA*
[4]*Memorial Health System, Colorado Springs, CO, USA*

*Corresponding author at: Department of Psychology, University of Colorado at Colorado Springs, 4047 Columbine Hall, Colorado Springs, CO 80918, USA.
Tel.: +1-719-255-4135; Fax: +1-719-255-4166.
E-mail address: bgavett@uccs.edu (B.E. Gavett).

Accepted 3 November 2011

Abstract

The serial position effect shows that two interrelated cognitive processes underlie immediate recall of a supraspan word list. The current study used item response theory (IRT) methods to determine whether the serial position effect poses a threat to the construct validity of immediate list recall as a measure of verbal episodic memory. Archival data were obtained from a national sample of 4,212 volunteers aged 28–84 in the Midlife Development in the United States study. Telephone assessment yielded item-level data for a single immediate recall trial of the Rey Auditory Verbal Learning Test (RAVLT). Two parameter logistic IRT procedures were used to estimate item parameters and the Q_1 statistic was used to evaluate item fit. A two-dimensional model better fit the data than a unidimensional model, supporting the notion that list recall is influenced by two underlying cognitive processes. IRT analyses revealed that 4 of the 15 RAVLT items (1, 12, 14, and 15) were misfit ($p < .05$). Item characteristic curves for items 14 and 15 decreased monotonically, implying an inverse relationship between the ability level and the probability of recall. Elimination of the four misfit items provided better fit to the data and met necessary IRT assumptions. Performance on a supraspan list learning test is influenced by multiple cognitive abilities; failure to account for the serial position of words decreases the construct validity of the test as a measure of episodic memory and may provide misleading results. IRT methods can ameliorate these problems and improve construct validity.

Keywords: Learning and memory; Assessment; Statistical methods

Introduction

List learning tests are commonly used by neuropsychologists to estimate episodic memory abilities (Rabin, Barr, & Burton, 2005). In a common paradigm, lists with word lengths that exceed an individual's attention span (supraspan) are presented to an examinee for immediate and delayed free recall. This form of memory assessment has demonstrated high levels of sensitivity and specificity to disorders causing memory dysfunction, such as amnestic mild cognitive impairment and Alzheimer's disease (AD; Gavett et al., 2009), temporal lobe epilepsy (Grammaldo et al., 2006), and post-concussive syndrome (Bazarian et al., 1999).

One commonly reported phenomenon inherent in all supraspan free recall list learning tests is the serial position effect (Deese & Kaufman, 1957). This effect describes the pattern by which the probability of a word being recalled varies as a function of its position in the list. Items from the beginning and the end of a list are more likely to be recalled than items from the middle of a list. Over the last several decades, theoretical models have been put forth to explain this phenomenon. Atkinson and Shiffrin (1968) proposed that recall of items from the beginning of the list (primacy effect) was facilitated due to an increased opportunity for these words to be rehearsed, and thus encoded into more stable memory stores (also *see* Rundus, 1971). The

doi:10.1093/arclin/acr104

Atkinson and Shiffrin model also proposed that recall of items from the end of a list (recency effect) was facilitated because these items were still present in the rehearsal buffer at the time of recall (also *see* Vallar & Papagano, 1986). This model was elaborated upon by Craik and Lockhart (1972), who proposed that rather than simply through increased rehearsal, earlier list items were remembered better because of the increased opportunity for the application of deeper encoding strategies. Later, Baddeley (1986) proposed a multi-component model of working memory, by which a central executive was responsible for acting upon information in low-level auditory and visual stores (the phonological loop and visuospatial sketch pad, respectively). Baddeley's (1986) model of working memory proposed that, in order for information to be encoded into stable memory stores, it must be acted upon by the central executive. These abbreviated descriptions of some of the most influential theoretical models of human memory simply serve to illustrate the fact that immediately recalling a supraspan list of words is thought to be controlled by multiple cognitive processes. For the sake of consistency with naming conventions in clinical neuropsychology, we have chosen to use the terms attention and short-term memory to describe primary and secondary memory processes, respectively.

Short-term memory and attention are dissociable constructs of human cognition. For instance, AD causes considerable damage to the neurobiological substrate of episodic memory, the hippocampal–entorhinal complex, early in the course of the disease (Sperling et al., 2010). Consistent with these known neuropathological changes, the clinical manifestations of AD show a disruption in episodic memory (Sperling et al., 2010) with general sparing of basic verbal attention early in the course of the disease when diagnosis is usually most valuable (Linn et al., 1995). This pattern of cognitive impairment caused by AD can be seen on supraspan list learning tasks. Patients with AD show a recency effect similar to cognitively healthy participants—suggesting intact attention—but show a reduced primacy effect consistent with short-term memory impairment (Gainotti & Marra, 1994; Spinnler, Della Sala, Bandera, & Baddeley, 1988). Despite the evidence for this dissociation between episodic memory and attention span in patients with AD, list learning test performance is nearly always interpreted based on the total number of words recalled from the list, regardless of the position of the word in the list. Unfortunately, a total sum does not identify the specific contributions made by attention versus short-term memory to the total score. For this reason, there is cause to suspect that typical neuropsychological approaches to list learning test interpretation lack construct validity (Cronbach & Meehl, 1955) for making unbiased estimates of episodic memory ability.

To address the potential influence of the serial position effect on immediate list recall, Buschke and colleagues (2006) examined the use of a scoring system that assigned weights to words recalled based on their position in the list. These authors hypothesized that, compared with unit weighting, assigning greater weight to words recalled from short-term memory stores (primacy effect) and lesser weight to words recalled from basic attention stores (recency effect) would more accurately distinguish individuals with mild AD from those without AD. The results were as expected: the AD group showed recency but not primacy effects and the weighted list learning score (area under the receiver operating characteristic curve [AUC] = 0.86) offered better discrimination of mild AD from controls than the unweighted score (AUC = 0.77).

Buschke and colleagues (2006) showed that the clinical detection of AD could be improved upon by accounting for the serial position effect in list learning test scoring and interpretation. However, the Buschke and colleagues (2006) study was limited by the use of theoretically derived item position weightings instead of empirically derived weightings. The serial position effect clearly shows that items from the middle of the list are the most difficult to remember, and therefore, an argument could be made that these items should be assigned the most weight. Fortunately, there are empirical methods to evaluate item-level test data for the purposes of deriving ability estimates.

Item response theory (IRT) is considered a part of modern psychometric theory, but its origins date back to the 1950s. Although used extensively in standardized educational assessment (e.g., the Graduate Record Examination), IRT has not factored largely into the development and validation of neuropsychological assessment instruments (but *see* Mungas, Reed, & Kramer, 2003; Mungas et al., 2010). IRT is based on the theory that each individual item that makes up a test contributes in some quantifiable way to the estimation of an underlying unidimensional trait (although multidimensional IRT is possible, it is rarely used). Depending on the model chosen, each item can be described by one (difficulty [β]), two (difficulty and discrimination [α]), or three (difficulty, discrimination, and guessing [γ]) parameters. These parameters allow for the estimation of an examinee's underlying ability (θ) and the quantification of a test's reliability and measurement error (Embretson & Reise, 2000) across the spectrum of ability levels. IRT differs from classical test theory (CTT) in that IRT requires stronger assumptions to be met with regard to the underlying latent trait and the items that contribute to its measurement (Lord & Novick, 1968; Nunnally & Bernstein, 1994). When applied to neuropsychological test construction and validation, IRT can be used to provide converging evidence for construct validity (Embretson & Gorin, 2001), to identify items that are not appropriate measures of the intended latent variable, to create test summary scores that weight items based on parameters such as difficulty, and to quantify the test's measurement error across all relevant ability levels (Mungas & Reed, 2000). The similarities and differences between IRT and CTT and ways that two methods augment one another have been addressed in great detail elsewhere (e.g., Embretson & Reise, 2000; Zickar & Broadfoot, 2009). Therefore, we will highlight the major advantages of IRT that are relevant in the current study.

In CTT, items on a particular test are usually assigned unit weights and summed to produce a total test score. In neuropsychological applications, that test score is then scaled relative to a normative sample to produce an estimate of an individual's ability on the trait assumed to be measured by the test. In IRT, tests can be scored using procedures that weight items based on the parameter estimates of the chosen model. Items that are low in information (a function of the chosen parameters and a similar construct to reliability) may be discarded or re-written to yield more desirable information, which serves to improve the test's standard error of measurement. Thus, IRT allows for an understanding of how each item contributes to the measurement of the underlying construct and to the test's measurement precision. In the case of list learning tests, if it is assumed that the test is intended to measure episodic memory, each item's contribution toward measuring this trait can be evaluated and items that do not appear valuable can be discarded. If lower-information items are not discarded, they are not weighted as heavily as items that possess higher information when deriving ability estimates based on the underlying construct. While IRT may be beneficial for enhancing the construct validity of list recall, the serial position effect, by definition, poses a challenge for IRT implementation. In addition to unidimensionality, IRT models also assume local independence; that is, the idea that an examinee's response to an item is based solely on his or her underlying ability level and not some other factor. In the case of list learning performance, the serial position effect clearly demonstrates that item position influences probability of recall, and thus, the local independence assumption appears to be violated. This will be a topic of investigation in the current study.

Based on the list learning research summarized above, it is believed that, consistent with the proposed mechanisms underlying the serial position effect, the number of words immediately recalled from a list is influenced by two related cognitive ability constructs: attention and short-term episodic memory. If true, this raises questions about the contributions of later items in the list to the test's construct validity as a measure of episodic memory. We hypothesize that a two-factor model of list learning performance will more closely approximate the item-level list learning data than a unidimensional model. We also predict that an IRT model can help to identify items that are more affected by attention than memory and that by eliminating these confounding items, the test's construct validity as a measure of memory will be improved. More specifically, we hypothesize that performance on later items in the list of words is largely a function of attention and therefore contributes little information about short-term memory abilities. Ignoring these items is not expected to change the measurement properties of the test under a unidimensional model of memory. Our goal in the current paper is to test these hypotheses in a sample of cognitively healthy adults and to establish the test's psychometric properties under an IRT model. This approach can help to determine whether the test has acceptable construct validity and measurement invariance for future research in clinical samples.

Method

Participants and Procedure

Participants were volunteers in the Midlife Development in the United States-II (MIDUS-II; Ryff & Lachman, 2007) study, a follow-up to a national survey of non-institutionalized adults selected by random-digit dialing (Brim, Ryff, & Kessler, 2004). Part of the MIDUS-II, completed between 2004 and 2006, included a computer-assisted telephone cognitive assessment using the Brief Test of Adult Cognition by Telephone (Tun & Lachman, 2006) in a sample of 4,212 participants. MIDUS-II Cognitive functioning archival data are publicly available through the Inter-University Consortium for Political and Social Research website (http://dx.doi.org/10.3886/ICPSR04652). Data were examined for missingness and other problems, as flagged by the MIDUS-II researchers. Cases that were flagged as problematic (for any cognitive test) or had missing data on the immediate recall trial of the list learning test were excluded. We extracted item-level responses from the immediate recall trial of the list learning test. Data were recoded in a binary fashion to indicate whether each item in the 15-word list was recalled by the participant (0 = not recalled, 1 = correctly recalled).

Measures

The list learning test administered to participants in the MIDUS study is a telephone version of the Rey Auditory Verbal Learning Test (RAVLT; Rey, 1964; Taylor, 1959), consisting of one immediate recall trial and one delayed recall trial. After checking for adequate hearing, the 15 list items were read to the participants with a 1-s pause between each word. The participants were then given 90 s to freely recall as many words as possible from the list, in any order. Correct responses, intrusions, and repetitions were recorded, but only correct responses on the immediate recall trial were used for the present study.

Statistical Analysis

Data analysis was conducted using R version 2.11.1 (R Development Core Team, 2010), including the ltm package for IRT analyses (Rizopoulos, 2006). For all significance tests, we set $\alpha = 0.05$. We calculated the mean score for each of the 15 items and regressed these scores onto the item sequence number (1–15) using a quadratic term to examine the serial position effect. The unidimensionality assumption of IRT was tested using modified parallel analysis (Drasgow & Lissak, 1983), a technique that employs Monte Carlo simulation to compare the matrix of tetrachoric correlations observed in the actual data under a specified IRT model to the simulated data (averaged across 200 simulations) under the same IRT model with unidimensionality assumed. We examined eigenvalues from the resulting scree plots and compared our results with the null hypothesis of no difference between the second eigenvalues of the observed data and the simulated data. We also conducted a visual examination of the scree plots in order to compare them with the exemplars depicted in Drasgow and Lissak (1983) for a determination of the robustness of the IRT model to potential violations of the unidimensionality assumption.

For all IRT analyses, we selected a two parameter logistic (2PL) model (*see*, e.g., Edwards, 2009), which estimates the difficulty (β) and discrimination (α) parameters of each item. These parameter estimates were used to create item characteristic curves, item information curves, test information curves, and to calculate the test's standard error of measurement. For all IRT analyses, we focused on estimating ability levels (θ) ranging from 3 *SD* above to 3 *SD* below the mean of an assumed normal distribution ($M = 0.0$, $SD = 1.0$).

We examined item fit using the Q_1 statistic (Yen, 1981), which approximates a χ^2 distribution. Due to the very large size of our sample and associated statistical power, it is likely that most, if not all, item fit statistics will be found to either under or over fit the model to a statistically significant degree (Reise, 1990). Therefore, instead of relying on an assumed χ^2 distribution to test for significant item misfit, we compared the real data with a 100-trial Monte Carlo simulation of the data under the null hypothesis (Yen, 1981). Significant differences in the Q_1 statistic between the real data and the simulated data suggest item misfit.

Because strong associations between item discrimination and item position suggest that the local independence assumption has been violated, we calculated the correlations among item position, item discrimination, and squared item discrimination to check for violations of this assumption (Reise & Waller, 1990).

Results

Of the 4,212 participants with data in the MIDUS-II Cognitive functioning data set, 239 were excluded due to missing data on the immediate word list learning task or due to questionable test validity on any portion of the cognitive evaluation. The remaining 3,973 participants ranged in age from 28 to 84 ($M = 55.8$, $SD = 12.3$) and included 2,133 (53.6%) women and 1,840 (46.3%) men (gender status was missing for five participants).

We found evidence for the serial position effect in the data, as can be seen in Fig. 1, which illustrates the accuracy obtained for each item in the current sample. The U-shaped trend line in Fig. 1 depicts the quadratic function that best fits these item accuracy data. The list learning data in the current sample show clear evidence of the serial position effect.

Having established the presence of a serial position effect in the current data, we tested the hypothesis that follows from this effect: that performance on a single 15-item word list recall trial is better explained by a model that accounts for two ability constructs (i.e., short-term episodic memory and attention) rather than a model that accounts for only one (i.e., short-term episodic memory). Using modified parallel analysis procedures (Drasgow & Lissak, 1983), a scree plot derived from the current data was compared with a Monte Carlo simulation of the data under an assumed unidimensional model (Fig. 2). As can be seen in Fig. 2, there are two eigenvalues >1.0. When compared with the simulated unidimensional data, the second eigenvalue of the observed data (1.19) is significantly larger than the second eigenvalue of the simulated data (0.31), $p =$ .005. We also compared the model fit of the actual data under a two-factor model with the model fit of the actual data under a unidimensional model. The results suggested that the two-dimensional model offered a significantly better fit to the data than the unidimensional model, D ($df = 15$) = 516.2, $p < .001$, with the second factor accounting for 8.2% of the total variance. Although these results provide evidence to suggest that the current list learning data are affected by two underlying cognitive constructs, the violation of the unidimensionality assumption is not so severe as to invalidate the use of a unidimensional IRT model according to guidelines set forth by Drasgow and Lissak (1983).

A 2PL IRT model yielded the item characteristic (top row) and item information curves (bottom row) shown in Fig. 3. Of particular salience, in the top left panel of Fig. 3, is the negative slope of the item characteristic curves for items 14 and 15,

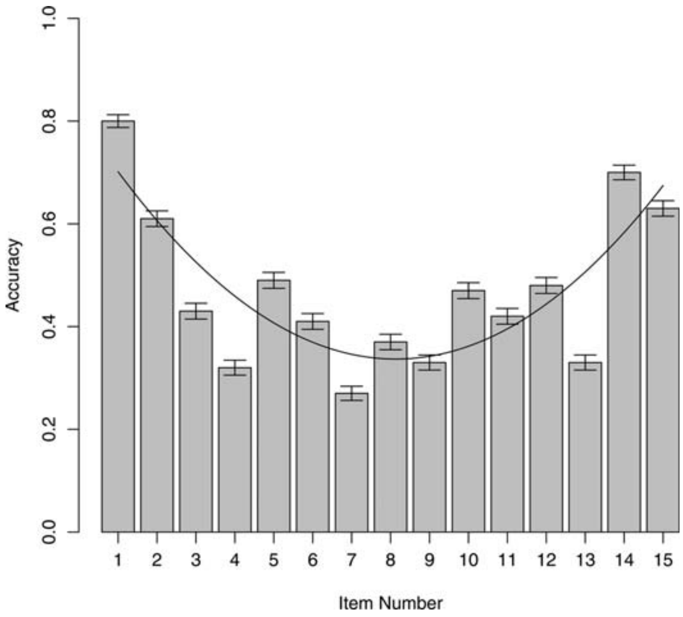

Fig. 1. The serial position effect in the current list learning data. The trendline represents the quadratic function that best fits the data. Error bars represent 95% confidence intervals.

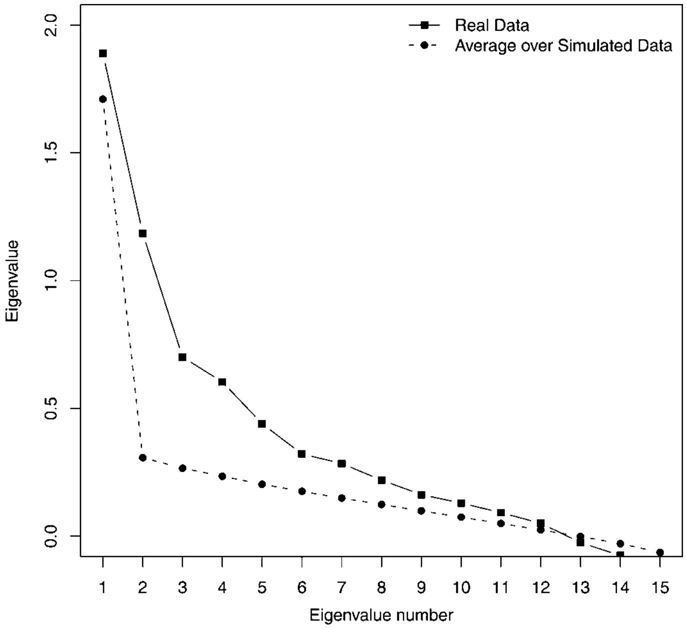

Fig. 2. Scree plot depicting the results of the modified parallel analysis for the 15-item list. The real data (solid line) show evidence of a two-factor solution based on the second eigenvalue >1.0 and the difference between the second eigenvalues of the real data the simulated data (dashed line).

which indicates that the lower an individual's memory ability, the more likely the person is to correctly recall these items; this effect is especially pronounced in the final item. Based on these data, 75% of individuals with episodic memory abilities falling 3 *SD* below the average will recall the final two words of this 15-item list. In contrast, items 14 and 15 are the least likely to be recalled by individuals with extremely high episodic memory abilities (*z*-scores ≥ 3.0). This pattern was replicated when the data were re-analyzed using a random-halving of the sample (data not shown). As a result of this undesirable pattern, these two items were excluded from further IRT analyses, which is a common approach to test development and refinement (Embretson & Reise, 2000). Item fit statistics, shown in Table 1, revealed significant item misfit for items 1, 12, and 15. The item

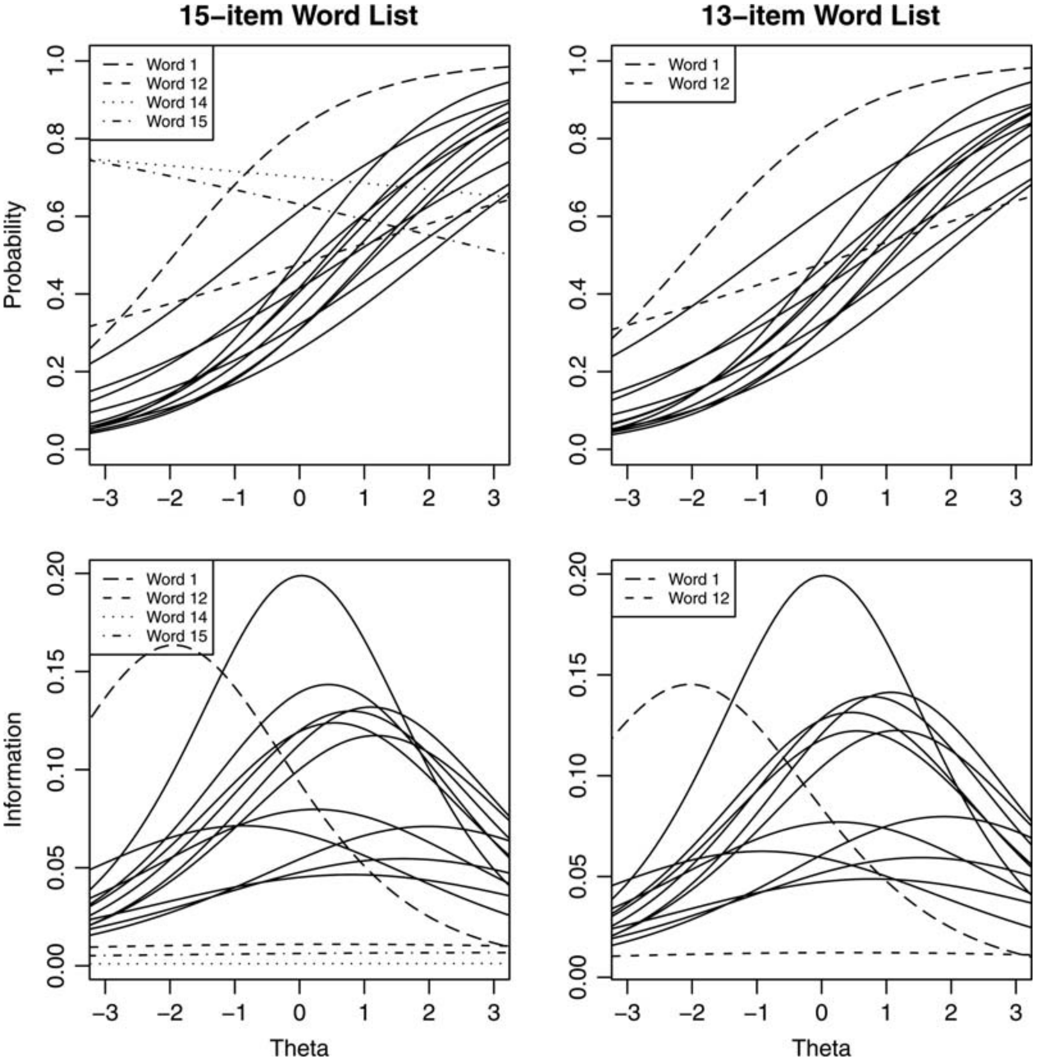

Fig. 3. Item characteristic (top row) and item information (bottom row) curves for the 15-item (left column) and 13-item (right column) 2PL IRT models. In the 15-item model, items 1, 12, 14, and 15 (broken lines) function dissimilarly from the remaining items (solid lines). Removing items 14 and 15 to yield a 13-item list did not improve the fit of items 1 and 12.

characteristic and information curves for these items and item 14 are highlighted in Fig. 3. In Figs 3–6, the *x*-axis represents the underlying ability (θ) measured by the test items, scaled as a *z*-score ($M = 0.0$, $SD = 1.0$), such that a θ value of zero represents an average ability level and higher θ *z*-scores represent higher ability estimates. The parameter estimates for the 15-item version of the RAVLT are likely biased by a violation of the local independence assumption; the correlations between item position and item discrimination were quite high ($r_{\mathrm{IP.a}} = -.79$; $r_{\mathrm{IP.a^2}} = -.76$).

Because the findings described above are consistent with our hypothesis, we sought to determine whether excluding the scores obtained on items 14 and 15 would change the measurement properties of the test, the fit of items 1 and 12, and the position-discrimination correlations. The item characteristic and item information curves from the 13-item model are presented on the right side of Fig. 3. The difficulty and discrimination parameters of the 13 items in this model are essentially unchanged compared with the original 15-item model (Table 1). We also derived the total test information functions for both the 13- and 15-item lists from the sum of the item information curves and used these to calculate the standard error of measurement for both versions of the test across the range of ability levels (Fig. 4). The elimination of the final two items from the 15-item test had no effect on the test's information or measurement precision. However, an examination of the item fit statistics again revealed significant misfit for items 1 and 12 (Table 1). Furthermore, the correlations between item position and item discrimination remained high ($r_{\mathrm{IP.a}} = -.58$; $r_{\mathrm{IP.a^2}} = -.56$), suggesting a violation of the local independence assumption. Therefore, a third model was tested, excluding items 1, 12, 14, and 15.

Table 1. IRT parameter estimates and item fit statistics for the 15- and 13-item models

Item number	Stimulus	15-item model				13-item model			
		β	α	Q_1	p-value	β	α	Q_1	p-value
1	Drum	−1.94	0.81	242.2*	.03	−2.03	0.76	214.6*	.05
2	Curtain	−0.87	0.53	128.4	.25	−0.92	0.50	108.4	.29
3	Bell	0.44	0.76	224.5	.17	0.46	0.72	211.4	.20
4	Coffee	1.18	0.69	156.9	.46	1.16	0.70	165.6	.41
5	School	0.03	0.89	274.1	.49	0.03	0.89	295.1	.29
6	Parent	0.54	0.70	183.6	.35	0.55	0.70	182.3	.42
7	Moon	1.99	0.53	107.6	.25	1.89	0.56	115.9	.23
8	Garden	0.79	0.72	192.1	.32	0.77	0.75	196.2	.42
9	Hat	1.10	0.73	170.3	.51	1.07	0.75	187.2	.50
10	Farmer	0.24	0.56	118.8	.60	0.24	0.56	113.4	.67
11	Nose	0.80	0.43	77.9	.47	0.79	0.44	86.8	.37
12	Turkey	0.44	0.21	50.0*	.02	0.41	0.22	62.0*	.04
13	Color	1.60	0.47	98.7	.17	1.54	0.49	103.2	.19
14	House	11.75	−0.07	18.2	.09				
15	River	3.27	−0.16	36.9*	.03				

Note: IRT = item response theory.
*$p < .05$.

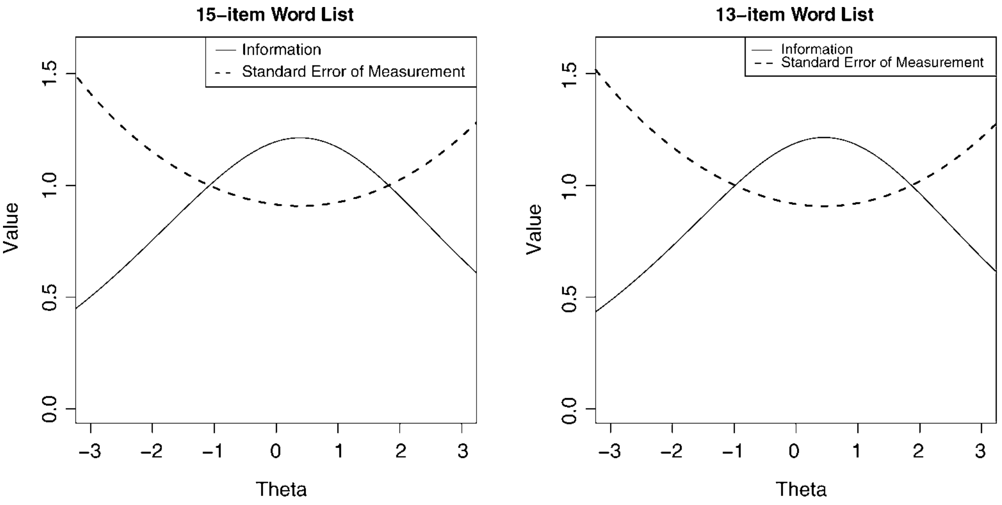

Fig. 4. Test information and standard error of measurement for the 15-item (left) and 13-item (right) 2PL IRT models. The similarity between the curves suggests that the measurement properties of the 15-item model are not impacted by the removal of the final two items.

Item fit analyses for the 11-item model revealed no significant misfit fit for any of the 11 items (Table 2). The exclusion of items 1, 12, 14, and 15 also kept the local independence assumption from being violated by attenuating the position-discrimination correlations ($r_{\mathrm{IP.a}} = -.05$; $r_{\mathrm{IP.a^2}} = -.08$). Item characteristic curves and item information curves for the 11-item list are displayed in Fig. 5. The 11-item test information function is plotted along with the 11-item standard error of measurement in Fig. 6.

Discussion

The serial position effect is one of the most well-established phenomena in the study of human memory. There is a large body of literature to suggest that immediate recall of a supraspan list of words is affected by two interrelated cognitive abilities, which we refer to here as attention and short-term episodic memory. Although this effect is known to most psychologists, neuropsychological test batteries regularly include tests of supraspan list recall as a measure of episodic memory without taking into account the possible influence of the serial position effect. The current results caution against the sole reliance upon a simple sum of items recalled, as failing to account for the position of words in a list reduces the construct validity of list recall as a measure of episodic memory. As expected, the final two list items were found to be inappropriate in a

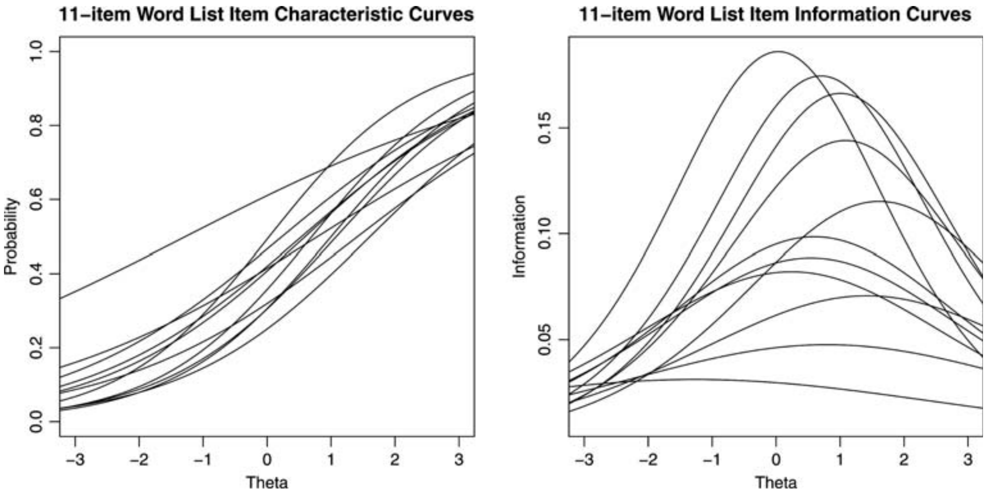

Fig. 5. Item characteristic (left) and item information (right) curves for the 11-item 2PL IRT model.

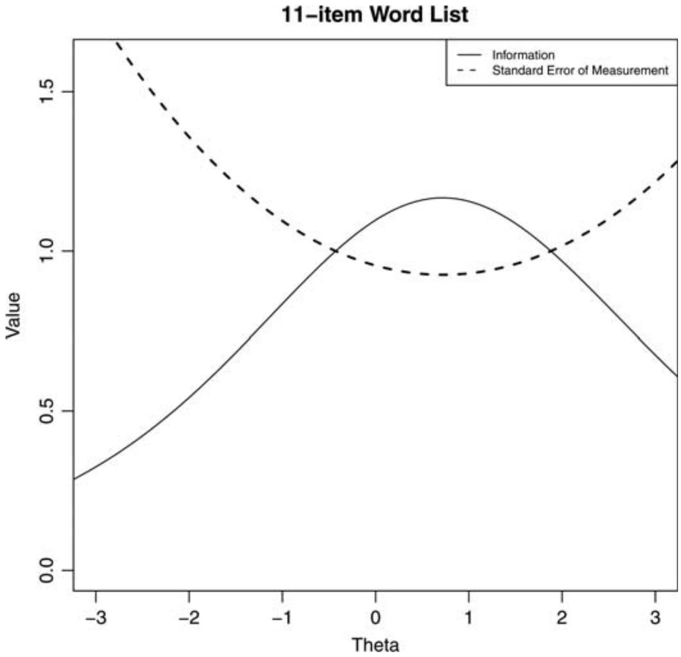

Fig. 6. Test information and standard error of measurement for the 11-item 2PL IRT model.

unidimensional model of episodic memory ability. The error introduced by these items is egregious; individuals with very low ability estimates are more likely to recall the last two items than individuals with very high ability estimates (Fig. 3). However, we also found evidence of misfit for items 1 and 12. Because item 1 is strongly associated with the primacy effect, its presence contributes to the violation of the local independence assumption, which precludes accurate parameter estimation under an IRT model. As the Atkinson and Shiffrin (1968) model suggests, the probability of recalling item 1 may be more strongly related to its position than to the examinee's underlying episodic memory skills. This may also be the case for item 12 with respect to the recency effect, but this explanation does not clarify why item 13, which should be expected to produce even stronger recency effects, was found to fit well within the model.

These findings are generally consistent with those reported by Buschke and colleagues (2006), who found that a weighting scheme that assigned increasingly greater credit to earlier items was more effective at detecting episodic memory impairment caused by AD than scoring with unit weighting. Although Buschke and colleagues (2006) correctly diverted points away from

Table 2. IRT parameter estimates and item fit statistics for the 11-item model

Item number	Stimulus	β (95% CI)	α (95% CI)	Q_1	p-value
2	Curtain	−1.27 (−1.66 to −0.88)	0.35 (0.25–0.45)	64.4	.34
3	Bell	0.54 (0.40–0.69)	0.60 (0.48–0.71)	172.7	.14
4	Coffee	1.08 (0.91–1.26)	0.76 (0.64–0.88)	194.3	.56
5	School	0.03 (−0.05–0.12)	0.86 (0.72–1.00)	294.0	.42
6	Parent	0.60 (0.45–0.74)	0.63 (0.51–0.75)	158.0	.50
7	Moon	1.61 (1.35–1.88)	0.68 (0.56–0.80)	139.0	.71
8	Garden	0.70 (0.58–0.82)	0.84 (0.70–0.97)	241.5	.60
9	Hat	1.00 (0.85–1.16)	0.82 (0.68–0.95)	226.4	.43
10	Farmer	0.24 (0.11–0.36)	0.57 (0.46–0.68)	148.2	.42
11	Nose	0.79 (0.57–1.02)	0.44 (0.34–0.54)	121.4	.16
13	Color	1.42 (1.13–1.71)	0.53 (0.42–0.64)	123.1	.24

Notes: IRT = item response theory; CI = confidence interval.

the final list items, the current findings suggest that assigning the greatest amount of weight to the earliest items may be a suboptimal approach to item weighting, especially considering that item 1, which receives the most weight under the Buschke and colleagues (2006) scoring system, did not fit our model of episodic memory.

The current results are also consistent with other research investigating the serial position effect in clinical samples. Patients with neurological illnesses that impair short-term memory through damage to the medial temporal lobes (e.g., AD) have a markedly reduced primacy effect with a relatively small change in recency effect during supraspan list recall (Gainotti & Marra, 1994; Spinnler et al., 1988). Based on our results, the lower an individual's memory ability, the more likely he or she is to recall items 14 and 15. Assigning equal weight to items 14 and 15, or greater weight to item 1 (as suggested by Buschke et al., 2006), may impede accurate interpretation of episodic memory test results, which could influence diagnosis and possible treatments.

The current results illustrate some of the benefits of using both IRT and CTT methods for neuropsychological test construction and validation and show why it is necessary for more neuropsychological instruments to be calibrated using IRT models. In the meantime, the utility of list learning tests may be improved through re-norming using methods that do not assign points for correct responses to misfit items. Without further IRT analyses or re-norming, accurate interpretation of list learning performance may be facilitated by close examination of the primacy and recency effect produced by test-takers. The Second Edition of the California Verbal Learning Test (Delis, Kaplan, Kramer, & Ober, 2000), for example, offers users the option of deriving normatively adjusted values for primacy and recency effects. Valuable clinical insight into an examinee's episodic memory abilities may be facilitated by the use of a process approach to list learning interpretation; that is, with consideration of the serial position and corresponding difficulty of individual items rather than simply interpreting the total number of words recalled.

Despite the information provided by IRT about the relative contribution of items in the derivation of latent ability estimates, IRT is not a psychometric panacea. For the measurement of short-term episodic memory, the IRT-derived 11-item list has better construct validity than the 15-item list, but nevertheless, it does not possess desirable measurement properties. The test information and the standard error of measurement functions (Fig. 6) show that the 11-item list is most appropriate for estimating abilities that fall between z-scores of -0.5 and 2.0; however, at these levels of θ, the standard error of measurement is approximately 1.0. In other words, the widths of the error bars for the 95% confidence intervals around θ are approximately 2.0 *SD* in both the positive and the negative directions. This imprecision suggests that a single trial immediate list recall test does not provide clinically useful information.

The current study has several limitations that may affect its applicability to clinical situations. First, the test data were collected via telephone evaluations and may not generalize to in-person testing scenarios. In addition, the methods used to collect the current data were not specifically designed to address the current hypotheses. Instead of randomizing the order of the 15 words, the same list order was administered to all participants. It is possible that the item parameters estimated in the current IRT model may be affected if some of the words in the list are inherently easier to remember than other words. For example, differences in word frequency (e.g., Thorndike & Lorge, 1944) may cause some words to be inherently more easy to remember than other words (Roodenrys & Miller, 2008); it is possible that this may account, in part, for the unexpected findings related to items 12 and 13. However, because evidence for the serial position effect was found, it is unlikely that these factors were substantial confounders. It is possible, however, that the somewhat unexpected findings related to the fit of items 12 and 13 are an artifact of the current sample. Although a random-halving of the current data yielded results consistent with the findings in the full sample, replication using an independent sample can help to ensure that the current findings are not an artifact of sampling anomalies. The unusual finding relating to items 12 and 13 may also suggest that the parameter estimates have been impacted by a small degree of local dependence, despite the fact that the removal of the four items seemingly reduced local dependence by a significant degree. Another potential limitation of this study is that participants were administered a single-trial list learning task, which is not the most common paradigm in clinical neuropsychology. Instead, most list learning tests present multiple immediate recall trials. It is unclear whether the problems related to construct validity can be attenuated by the use of scores that summarize multiple recall trials instead of a single trial. Extension of the current methods to include in-person, multitrial list learning tests is likely to yield results that are more clinically meaningful and, ideally, more precise. Finally, these results are based on a sample of cognitively healthy adults, so external validation in a known clinical group was not possible using the MIDUS-II data. The extent to which these results are clinically meaningful will depend in part upon replication and extension in a memory disordered sample (e.g., AD). However, valid assessment of memory and other constructs in clinical samples is dependent on construct validity and the assumption of measurement invariance (Borsboom, 2006; Meredith & Teresi, 2006); therefore, the current results point to areas in which the validity of memory measurement in clinical samples may be improved upon. Future studies should seek to determine whether there is evidence of item- and test-level measurement invariance based on clinical and other groupings.

In conclusion, we tested the hypothesis that a single supraspan list learning test is affected by both memory and attention, as suggested by research into the serial position effect. If true, the contributions of attention and memory to the total test score cannot be separated by traditional scoring methods (i.e., total score based on unit weighting); as a result, the recency effect may lead to artificially inflated test scores, especially in individuals with low episodic memory ability (e.g., caused by AD). The current results suggest that removing four items from the RAVLT, under an IRT model, does not reduce the psychometric properties of the test as a measure of latent memory skills, but does reduce the confounding effects of attention on this test. These findings suggest a potential method for improving the assessment of episodic memory ability in adults; applying these findings to clinical groups should be a focus of future research.

Funding

This work was supported by the National Institute on Aging (grant number P30AG13846 and supplement 0572063345-5); the National Operating Committee on Standards for Athletic Equipment; and by an unrestricted gift from the National Football League. The funding sources were not involved in the preparation, review, or approval of this article.

Conflict of Interest

None declared.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York: Academic Press.

Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press.

Bazarian, J. J., Wong, T., Harris, M., Leahey, N., Mookerjee, S., & Dombovy, M. (1999). Epidemiology and predictors of post-concussive syndrome after minor head injury in an emergency population. *Brain Injury*, *13*, 173–189.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, S176–S181.

Brim, O., Ryff, C., & Kessler, R. (2004). *How healthy are we?: A national study of well-being at midlife*. Chicago: University of Chicago Press.

Buschke, H., Sliwinski, M. J., Kuslansky, G., Katz, M., Verghese, J., & Lipton, R. B. (2006). Retention weighted recall improves discrimination of Alzheimer's disease. *Journal of the International Neuropsychological Society*, *12*, 436–440.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Deese, J., & Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*, 180–187.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *The California Verbal Learning Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Drasgow, F., & Lissak, R. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363–373.

Edwards, M. C. (2009). An introduction to item response theory using the Need for Cognition Scale. *Social and Personality Psychology Compass*, *3/4*, 507–529.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gainotti, G., & Marra, C. (1994). Some aspects of memory disorders clearly distinguish dementia of the Alzheimer's type from depressive pseudo-dementia. *Journal of Clinical and Experimental Neuropsychology*, *16*, 65–78.

Gavett, B. E., Poon, S. J., Ozonoff, A., Jefferson, A. L., Nair, A. K., Green, R. C., et al. (2009). Diagnostic utility of the NAB List Learning test in Alzheimer's disease and amnestic mild cognitive impairment. *Journal of the International Neuropsychological Society*, *15*, 121–129.

Grammaldo, L. G., Giampa, T., Quataro, P. P., Picardi, A., Mascia, A., Sparano, A., et al. (2006). Lateralizing value of memory tests in drug-resistant temporal lobe epilepsy. *European Journal of Neurology*, *13*, 371–376.

Linn, R. T., Wolf, P. A., Bachman, D. L., Knoefel, J. E., Cobb, J. L., Belanger, A. J., et al. (1995). The 'preclinical phase' of probable Alzheimer's disease. A 13-year prospective study of the Framingham cohort. *Archives of Neurology*, *52*, 485–490.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69–S77.

Mungas, D., Beckett, L., Harvey, D., Farias, S., Reed, B., Carmichael, O., et al. (2010). Heterogeneity of cognitive trajectories in diverse older persons. *Psychology and Aging*, *25*, 606–619.

Mungas, D., & Reed, B. R. (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine*, *19*, 1631–1644.

Mungas, D., Reed, B., & Kramer, J. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, *17*, 380–392.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

R Development Core Team. (2010). *R: A language and environment for statistical computing [computer software]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*, 33–65.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127–137.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Roodenrys, S., & Miller, L. M. (2008). A constrained Rasch model of trace redintegration in serial recall. *Memory and Cognition*, *36*, 578–587.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63–77.

Ryff, C. D., & Lachman, M. E. (2007). *National Survey of Midlife Development in the United States (MIDUS II): Cognitive Project, 2004–2006 [Data file]*. ICPSR25281-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 13-07-2010. Retrieved from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/04652. doi:10.3886/ICPSR25281.

Sperling, R. A., Dickerson, B. C., Pihlajamaki, M., Vannini, P., LaViolette, P. S., Vitolo, O. V., et al. (2010). Functional alterations in memory networks in early Alzheimer's disease. *Neuromolecular Medicine*, *12*, 27–43.

Spinnler, H., Della Sala, S., Bandera, R., & Baddeley, A. (1988). Dementia, ageing, and the structure of human memory. *Cognitive Neuropsychology*, *5*, 193–211.

Taylor, E. M. (1959). *Psychological appraisal of children with cerebral deficits*. Cambridge, MA: Harvard University Press.

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Tun, P. A., & Lachman, M. E. (2006). Telephone assessment of cognitive function in adulthood: The Brief Test of Adult Cognition by Telephone. *Age and Ageing*, *35*, 629–632.

Vallar, G., & Papagano, C. (1986). Phonological short-term store and the nature of the recency effect: Evidence from neuropsychology. *Brain and Cognition*, *5*, 428–442.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance, & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in organizational and social sciences* (pp. 37–59). New York: Routledge.