An Investigation of the Measurement Properties of the Spot-the-Word Test in a Community Sample

Andrew Mackinnon and Helen Christensen
The Australian National University

Intellectual ability is assessed with the Spot-the-Word (STW) test (A. Baddeley, H. Emslie, & I. Nimmo Smith, 1993) by asking respondents to identify a word in a word-nonword item pair. Results in moderate-sized samples suggest this ability is resistant to decline due to dementia. The authors used a 3-parameter item response theory model to investigate the measurement properties of the STW in a large community-dwelling sample ($n = 2,480$) 60 to 64 years of age. A number of poorly performing items were identified. Substantial guessing was present; however, the number of words correctly identified was found to be an accurate index of ability. Performance was moderately related to a number of tests of cognitive performance and was effectively unrelated to visual acuity and to physical or mental health status. The STW is a promising test of ability that, in the future, may be refined by the deletion or replacement of poorly functioning items.

Keywords: premorbid intelligence, dementia, item response theory, guessing, item and test information

Knowledge of an individual's cognitive performance earlier in life is important for those determining the extent to which these abilities have declined when the individual is diagnosed as having a dementing illness. The extent of deterioration may be obvious when a severe disorder is manifest; however, early in the progression of the disorder, it may be difficult or impossible to differentiate those who have suffered a decline in cognitive ability from those whose performance has been poor throughout life. These problems are magnified when less severe impairments, such as mild cognitive disorder or age-associated memory impairment (see Christensen et al., 1995), are assessed. In these situations, high premorbid performance may mask a decline in ability. The assessment of premorbid ability is also relevant when one is assessing brain injury or damage arising from accidents or other illnesses. In all of these situations, recourse to assessments made earlier in life would be the most desirable method of calculating change (Crawford, Stewart, Parker, Besson, & Cochrane, 1989). Such information is rarely available, and thus methods that estimate premorbid ability must be used.

Although several approaches to estimating premorbid IQ have been devised (Beardsall & Hupert, 1997; Crawford, 1989, 1990;

O'Carroll, 1995; Paolo & Ryan, 1992), two methods have the greatest support. The first relies on educational and occupational attainment and on socioeconomic status. This information is obviously robust to decline and has been shown to correlate surprisingly highly with actual ability in some settings (Crawford et al., 1989). However, there are substantial difficulties in calibration of this approach for different countries, sexes, and, presumably, age cohorts. Such an approach can classify individuals only relatively coarsely and cannot even approach the status of a parallel form of any test of ability.

The second approach to measuring premorbid performance is by the use of psychometric tests that tap ability in a manner robust to recent decline. The most widely used of such tests is the National Adult Reading Test (NART; Nelson & O'Connell, 1982). This test consists of 50 words, graded in difficulty, whose pronunciation cannot be determined from their spelling. The participant is asked to read these words aloud from a printed card and is scored on the number of words pronounced correctly. NART scores correlate highly with Wechsler Adult Intelligence Scale—Revised (WAIS-R; Wechsler, 1981) full-scale and verbal IQ scores. Nelson (1982) provided regression equations that yield estimates of WAIS scores from the NART.

Despite widespread use in clinical and research settings, the NART has a number of significant limitations. Because it relies on orthographic irregularity, this approach is difficult to adapt to languages other than English. Analogues of the NART have been produced in Dutch (Schmand, Bakker, Saan, & Louman, 1991) and in French (Mackinnon & Mulligan, 2005; Mackinnon, Ritchie, & Mulligan, 1999), both of which have much more regular pronunciation than does English. However, in many languages, this approach is not possible. Even within English, there is a diversity of acceptable pronunciation. Words pronounced irregularly in British English are regularized (or almost so) in American English. This linguistic pattern has given rise to adaptations of the NART to American English (Uttl, 2002) and to Australian English (Lucas,

Andrew Mackinnon and Helen Christensen, Centre for Mental Health Research, the Australian National University, Canberra, Australia.

Funding to collect the data used in this study was provided by National Health and Medical Research Council (NHMRC) Program Grant 179805. Helen Christensen is supported by NHMRC Fellowship 366781. We wish to thank Bryan Rodgers, Kaarin Anstey, Anthony Jorm, and Perminder Sachdev (chief investigators for Wave 1 of PATH) as well as Trish Jacomb, Karen Maxwell, and the PATH interviewers for their contribution to the study.

Correspondence concerning this article should be addressed to Andrew Mackinnon, Centre for Mental Health Research, The Australian National University, Canberra 0200, Australia. E-mail: Andrew.Mackinnon@anu.edu.au

Carstairs, & Shores, 2003). The recently released Wechsler Test of Adult Reading (Psychological Corporation, 2001), which is conceptually identical to the NART and also uses this approach, has the advantage of having been normed with the most recent editions of the WAIS and the Wechsler Memory Scale (Wechsler, 1997).

In many settings, an important limitation of the NART is that it must be administered on an individual basis and that the interviewer must be trained to judge whether a word has been pronounced correctly. These requirements severely limit use of the NART in community and epidemiological surveys, as well as in other situations in which data are collected by self-completion methods or in group settings. In clinical settings, pronunciation-based tests cannot be used with individuals who have dysphasia or articulatory problems.

To overcome these and other problems with the NART, Baddeley, Emslie, and Nimmo Smith (1993) developed a test based on the discrimination of words from nonwords or pseudo words. The Spot-the-Word (STW) test comprises 60 word-nonword pairs (for example, *scion*-*frebulant*). The words range from common to rare and obscure terms. The nonwords were constructed so as to be the same length as the real words and to follow regular spelling and pronunciation rules. Participants are required to identify the real word in each pair. The STW overcomes many limitations of the NART.

Although the literature on the STW does not mention it, the concept behind the test was predicated in German by the widely used Mehrfachwahl-Wortschatz-Intelligenztest B (MWT-B; Lehrl, 1971). This test requires respondents to identify the real word among three nonword alternatives. While principally used as a measure of current verbal intelligence, the MWT-B has been used as a measure of premorbid ability (Mackinnon & Mulligan, 2001). The MWT-B has been the subject of a great deal of validation research and development over a long period. This research has provided general evidence that supports the notion of using word-nonword discrimination as a test of intellectual ability.

Performance on the STW has been shown to be highly correlated with performance on other tests of vocabulary and fluid intelligence and also on the NART (Baddeley et al., 1993; Beardsall & Hupert, 1997; Crowell, Vanderploeg, Small, Graves, & Mortimer, 2002; Lucas et al., 2003; Yuspeh & Vanderploeg, 2000). It has been found to be resistant to minimal but not more advanced Alzheimer's disease (AD; Beardsall & Hupert, 1997; Law & O'Carroll, 1998). More recently, McFarlane, Welch, and Rodgers (2006) extended this finding by reporting comparable performance on the test across nondemented, minimal AD, and mild AD groups, a finding that supports the STW as a relatively robust measure of premorbid ability. In younger samples, the STW has been found to be resistant to the effects of brain injury (Hinton-Bayre, Geffen, & MacFarland, 1997; Watt & O'Carroll, 1999). However, any conclusions drawn from these findings must be tempered by their small samples and by a recruitment procedure that limits the generalizability of results.

Although the STW is being used increasingly and offers a number of potential advantages over other methods of assessing premorbid ability, there is a paucity of research into the measurement characteristics of the test. In the only population-based study of the test, Crowell et al. (2002) reported norms for the STW in a moderate-sized, community-based sample ($N = 394$). They also reported correlations of the STW with age and education compa-

rable to those found for the WAIS. Beyond estimates of reliability, little information is available about the behavior of the STW items, individually and collectively, in discriminating a dimension of ability. These dimensions include the efficiency of items and the level of ability at which they function, susceptibility to guessing, and the effect of using simple scoring schemes in the presence of guessing.

Investigation of the measurement properties of the STW requires an analytic framework that is suited to binary items and allows for the possibility of guessing. It also requires a large, representative sample, so investigators can examine performance over the full range of normal intellectual ability and can obtain stable and precise parameter estimates. Our aim in this study was to investigate the measurement properties of the STW in a large, older community sample. In the absence of any previous analyses of the STW, the discriminatory power of items was of primary interest. We employed a three-parameter item response theory (IRT) model to provide more information about item fit and test performance and to identify candidate items for removal or replacement. Given the potential for substantial guessing effects and the uncertain effect of guessing on the accuracy of using number correct as an estimate of performance, these scores were compared with ability estimates derived with IRT. We examined the relationship of each type of estimate with demographic status, educational attainment, cognitive performance, and physical and mental health to contribute to the assessment of the convergent validity of the test. Convergent validity would be indicated by positive relationships with education and cognitive ability and by absent or negligible relationships with health measures.

Three-Parameter IRT Model

In the IRT analyses presented here, the probability of correctly identifying a word and a level of ability is summarized by a three-parameter item characteristic curve (ICC),

$$P(x_{ij} = 1 | \theta_i) = c_i + (1 - c_i) \frac{1}{(1 + e^{-D a_j (\theta_i - b_j)})}, \quad (1)$$

where $x_{ij} = 1$ or $x_{ij} = 0$, respectively, when Respondent i responds correctly or incorrectly to Item (word-nonword pair) j . The three parameters for each item are discriminability, as indexed by the slope of the ICC (a_j); location, as characterized by the ICC threshold (b_j); and "guessing," indicated by the asymptote (c_j). The location of the respondent on the ability dimension is denoted by θ_i . The constant D is assigned the value 1.7, so that the logistic response function matches the normal ogive model as closely as possible. Fuller details on this model and on IRT are given by Lord (1980) and by Hambleton and Swaminathan (1985).

Effects of Guessing

Consideration of the appropriate treatment of guessing has a long history in the scoring of multiple-choice tests (Dunlap, De Melo, & Cureton, 1929). Various schemes have been proposed. These often involve "correcting" number right scores for guessing or penalizing incorrect answers. Such procedures are frequently applied in conjunction with instructions to test takers either to omit items when they are unsure of the correct answer or to answer all items regardless of any lack of confidence about the correct

alternative. There is little evidence to support the use of these approaches. IRT offers a defensible, empirically based approach to the treatment of guessing. It also reveals the futility of ad hoc schemes that purport to correct for guessing, as it demonstrates that any weighting, discounting, or correcting for guessing must be based on the ability of the individual concerned. When a challenged respondent answers a difficult item correctly, this is likely to be due to guessing. However, chance will not play such a role in the case of an able respondent. Thus, a priori weights or adjustments, applied without consideration of where a respondent lies on the continuum being assessed, cannot be defensible. The actual effects of guessing and of its treatment, either by ignoring it or by the use of misapplied schemes, are unclear. However, it is apparent that simplistic approaches have the potential to produce inaccurate estimates of ability.

Reliability and Precision of Measurement

In classical test theory, error of measurement is indexed by the reliability coefficient. This value is an average over the range of ability measured by the test. IRT allows the precision of measurement along the latent dimension to be assessed for individual items and for tests. Precision of measurement at a point θ on the ability dimension is referred to as information $I(\theta)$. The standard error of measurement is inversely related to information:

$$SE(\theta) = 1/\sqrt{I(\theta)}. \quad (2)$$

The maximum information¹ of an item and the point at which it is attained are given by

$$I(\theta)_{\max} = \frac{D^2 a_i^2}{8(1 - c_i)^2} [1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{3/2}] \quad (3)$$

and

$$\theta_{\max} = \frac{1}{Da_i} \ln[1/2 + 1/2 \sqrt{1 + 8c_i}]. \quad (4)$$

It can be seen from Equation 3 that information is an increasing function of an item's slope but a decreasing function of the guessing parameter. Given equal slopes, an item with no guessing ($c_i = 0$) offers over eight times the information of one with a 50% chance of guessing correctly. Total test information can be shown to be the sum of the information functions for each item. Some idea of desirable values of this measure can be gained by considering that if a test has information of 10 at a particular point, the corresponding standard error is .31, yielding a test reliability of .90 at that point.

Consideration of the contribution of an individual item to test information and of the location where it is effective is useful for test design and refinement. Items that contribute little information, that are effective in regions of ability where few respondents lie, or in which there is no interest in accurate measurement (e.g., measuring high ability when screening for cognitive impairment) are candidates for removal. When items work in the same region, effectively duplicating one another and producing a peak of information in one region, consideration can be given to eliminating one or more items. Thus, IRT provides much more information than do approaches based on factor analysis or classical test theory as a means of eliminating poorly

functioning items or of efficiently reducing test length with minimum impact on precision of measurement.

Method

Participants

The sample came from the oldest cohort of the PATH Through Life Project, a large community survey concerned with the health and well-being of people 20–24, 40–44, and 60–64 years of age who live in the Canberra and Queanbeyan region of Australia. Results presented here concern the first wave interviews with 60- to 64-year-olds, which were conducted in 2001 or early 2002. Participants had to be in this age group on January 1, 2001. The sampling frames were the electoral rolls for Canberra and Queanbeyan, Australia. Registration on the electoral roll is compulsory for Australian citizens. Letters were sent to 4,832 persons. Of these, 34 were out of the required age range, 182 had moved, 28 were dead, 209 could not be found, and 1,827 refused to participate or their English was too poor to allow an interview. Thus, 2,551 individuals were interviewed (58.3% of those found and in the age range). The analyses presented were conducted on 2,482 participants who answered all STW items. This group comprised 1,199 men and 1,283 women. At the time of testing, their mean age was 62.5 years ($SD = 1.5$). Participants had, on average, 13.85 years of education ($SD = 2.71$, range 4–19.5).

Instruments

The measures used in the current analyses were taken from a broader study of cognitive ability, personality, well-being, and physical and mental health. The constraints of a large survey precluded the comprehensive measurement of ability. We chose the measures included here to tap cognitive ability as well as possible in this context and to investigate possible extraneous influences on performance on the STW.

STW. As described above, this 60-item lexical decision-making test requires the participant to choose the correct word from a pair of words containing one word and one nonword. Version A was used, and the test was administered as described in the manual (Baddeley, Emslie, & Nimmo Smith, 1992).

Digit Span Backward. This is a subtest of the Wechsler Memory Scale (Wechsler, 1945). Participants were read a series of digits and had to recall the digits in reverse order. There were two trials of series ranging from three to seven digits. The test was discontinued when the participant made errors on two trials at the same level. The score was the total number of correct trials.

Immediate recall. Immediate recall was assessed using the first trial of the California Verbal Learning Test (Delis, Kramer, Kaplan, & Ober, 1987). After being read a list of 16 nouns, participants were asked to recall as many words as possible.

Delayed recall. A delayed recall test of the immediate recall items was administered after an interval of approximately 3 min.

Simple reaction time and complex reaction time. Reaction time was tested using a small box that the participant held with

¹ Note that the formula given by Hambleton and Swaminathan (1985) for $I(\theta)_{\max}$ (Equation 6.17, p. 105, and Table 6-1, pp. 106–07) contains a minor error.

both hands; left and right buttons, at the top, were depressed with the index fingers. The front of the box had three lights: two red stimulus lights under the left and right buttons, respectively, and a green get-ready light in the middle, beneath these. There were four blocks of 20 trials that measured simple reaction time, followed by two blocks of 20 trials that measured complex reaction time. Means were calculated after outliers were removed. This was done by first eliminating any values over 6,000 ms. Next, means and standard deviations were calculated for each block for each participant, and values lying outside three standard deviations were eliminated. Finally, means were calculated again. Means for remaining outliers were truncated at 1,000 ms.

Symbol Digit Modalities Test (SDMT; Smith, 1973). The participant was required to substitute digits (1 through 9) corresponding to symbols. Symbol-digit pairs to be used were available throughout the task on a code sheet. The participant was required to make substitutions as rapidly as possible. The final score was the number of substitutions completed correctly in 90 s.

Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975). The MMSE evaluates current cognitive status. Components of the test include orientation as to time and place, memory, arithmetic, language use and comprehension, and copying a drawing. Possible scores range from 0 to 30. The MMSE is often used as a screening test for dementia (Tombaugh & McIntyre, 1992).

Vision. Visual acuity was measured with a 3-m Snellen chart. It contained seven lines that subtended 1 min of an arc at distances of 60, 36, 24, 16, 9, 6, and 5 m. Corrected vision was used. Possible scores range from 0 to 28 items correctly identified.

Handgrip. Peak grip strength was measured using the Smedley hand dynamometer (Model No PE7, Stoelting Co., Wood Dale, Illinois), which measures the force exerted in kilograms. Four trials were given for each hand, and the mean of the average score of left and right hands was the measure used in this study.

Forced expiratory volume (FEV). Pulmonary function was indexed by FEV in 1 s using a "micro" spirometer (Micro Medical Limited, Rochester, Kent, UK), which is a handheld microprocessor-based unit. The average of the measures was used. FEV was adjusted by dividing by body mass index to control for height and weight. FEV is a primary measure of lung function used here as a general indicator of physical health.

Short Form 12 (SF-12). The SF-12 is an abbreviated form of a 36-item instrument (SF-36) designed to assess general self-rated health, physical and psychological symptoms, and limitations in everyday activity due to physical and mental health over the previous 4 weeks (Ware, Kosinski, & Keller, 1996). In the current sample, Cronbach's alpha and mean interitem correlations were .80 and .40 for the SF-12 Physical Health Scale and were .80 and .39 for the SF-12 Mental Health Scale.

Goldberg scales. These scales comprise two sets of nine items that measure depression and anxiety, respectively. Respondents answer questions concerning "how you have been feeling in the past month." The scales have been validated for detecting elevated levels of depression and anxiety in epidemiological studies (Mackinnon et al., 1994). In the current sample, Cronbach's alpha and mean interitem correlations were .78 and .28 for the Goldberg Depression Scale and were .76 and .26 for the Goldberg Anxiety Scale.

Procedure

Participants completed a questionnaire on handheld computers. The topics covered included sociodemographic characteristics, anxiety and depression, substance abuse, cognitive function, well-being, physical health, health habits, use of health services, personality, coping, early life psychosocial risk factors, current psychosocial risk factors, and nutrition. The questionnaire was completed under the supervision of a professional interviewer. Tests of physical functioning—including blood pressure, grip strength, visual acuity, and lung function—were carried out. Most of the interview was self completed on a Hewlett-Packard 620LX palmtop personal computer using the Surveycraft software for computer-assisted personal interviewing. However, testing by the interviewer was required for the vision test and some of the cognitive tests.

Results

Classical Test Theory Statistics

Cronbach's alpha for the 60 items was .85 (95% confidence interval .87–.91).² The average interitem correlation was .12 (range = -.16–.47). For individual pairs, the percentage of words correctly identified ranged from 31.1% to 99.6%. In two pairs, rates of correct identification were substantially below theoretical chance levels. Identified by the word in the word-nonword pair, they were *threnody* (31.1%) and *narwhal* (40.2%). Words in 10 pairs were correctly identified by over 99% of participants.

Dimensionality

A preliminary requirement for IRT analysis is that the scale formed by the items is essentially unidimensional. Unidimensionality was assessed by examination of the eigenstructure of the matrix of tetrachoric correlations between the items of each scale (Lord, 1980). The analysis showed one dominant eigenvalue accounting for 41.1% of the total variance. The ratio of the first two eigenvalues was 4.06 (23.31/5.74). Low and negative correlations between some items suggested that some individual items may not relate to the ability dimension under investigation, but there was no evidence of multidimensionality. Reckase (1979) found that parameter estimates for IRT models reflect the first principal component of the items under these circumstances. Thus, the parameters reported here relate to the full STW scale, as it is commonly used.

IRT Statistics

We used Bilog-MG Version 3.0 (Zimowski, Muraki, Mislevy, & Bock, 1996) to estimate the parameters in the logistic IRT models tested. Marginal maximum likelihood estimation was used. Except as follows, program defaults were retained. Criteria for convergence were set to the more rigorous value of 0.0001, and the number of iterations permitted was increased to achieve convergence. Prior distributions were specified for threshold, slope, and

² The confidence interval was calculated with an SPSS macro written by Marso (2007). For the interval, 5,000 bootstrap samples were used.

asymptote parameters. For asymptotes, we chose parameters for the beta distribution to reflect two alternatives. Prior distributions were imposed on thresholds because of the difficulty of estimating asymptotes for easy items. In the absence of prior distributions, thresholds of easy items tended to drift toward infinity as asymptote parameters approached zero. Under this configuration, convergence was achieved in 107 iterations.

No omnibus test of model fit is available for a test comprising this many items. However, it is possible to compare the fit of the three-parameter model to that of the simpler two-parameter model, which is effectively the former model with asymptotes constrained to be zero. The difference in $2 \times \text{Log Likelihood}$ between models is distributed as a chi-square, with degrees of freedom equal to the number of items. For the current data, this difference was $\chi^2(60, N = 2,482) = 1362.68, p < .05$, which indicates that the effects of guessing were substantial. Fitting a model in which the asymptote parameter was constrained to the same value for all items was also associated with a smaller but still substantial decline in fit, $\chi^2(59, N = 2,482) = 350.10, p < .05$, which indicates that guessing effects varied between items.

In addition to the above tests, BILOG provides a test of fit for each item (see Table 1). This is achieved by dividing the sample into a number of groups on the basis of estimated ability and comparing the number of participants expected to answer correctly under the IRT model to that actually observed. These deviations form the basis of a chi-square of the goodness of fit of a single item to the model. Test results are shown in Table 1. Due to the large sample size and the number of tests undertaken, a statistically significant result may arise from a trivial misfit. However, for a number of items, substantial discrepancies were observed. Figure 1 shows four misfitting items. All share the characteristic that persons with low ability were more likely to choose the correct response than were somewhat more able individuals. This is most apparent for Item 14 (*threnody*). For these word-nonword pairs, the nonword is unattractive to those of low ability; thus, these items cannot discriminate low ability from high ability. Note that only one item (14) had a negative item-total correlation, an index frequently used in classical test theory to identify poorly functioning items.

Threshold, slope, and asymptote parameters for each item are shown in Table 1. Slope may be transformed into a factor loading on the ability dimension. These values range from 0.34 to 0.97 ($M = 0.77$), a finding that indicates the majority of items have very substantial power to discriminate the underlying ability dimension. With the exception of Item 14 (the word pair that showed particularly poor fit), all slope and loading parameters were statistically significantly different from zero. As would be expected, given that choice is made from a pair, guessing parameters were quite high, ranging from 0.17 to the constrained upper boundary of 0.50 ($M = 0.38$). Discounting poorly fitting items, the lowest guessing parameter was 0.28.

Item and Test Information

The maximum information for each item and the location on the ability dimension of the maximum are shown in Table 1. Note that although Item 14 had the largest maximum information, the standard error of this quantity (6.71) was very large as well. As the slope was not significantly different from zero, this item cannot

provide any reliable information about the location of respondents. This finding applied to a lesser extent to the item with the second largest maximum (Item 56). Although significantly greater than zero, the standard error of the maximum (1.68) was quite high. For both items, the fit of the model to the data was quite poor, and thus the item parameters and indices derived from them are questionable.

The information of the test was calculated by summing over items at specific levels of ability. The resultant test information curve is shown in Figure 2. Also shown is the information curve that resulted from deletion of the two particularly poorly fitting items considered above (Items 14 and 56). As is typical, this curve is peaked, which shows that maximum precision in locating respondents on the dimension occurred approximately one standard deviation below the mean. Good precision was maintained down to three standard deviations below the mean. This precision was not mirrored at or above the mean, particularly when the two poorly fitting items that made a disproportionate contribution to test information were removed.

Item parameters and information may be used to refine tests in a variety of ways. Items with low information maxima may be removed with little impact on test precision. Where a number of items have maxima in the same region of ability, the redundant items that add the least information may be candidates for removal.

Estimation of Ability (θ)

Like item parameters, the locations of individuals on the trait are parameters to be estimated in the IRT model using programs such as BILOG. Because the estimation of participants' ability requires access to IRT software, it is of interest to know how well the conventional approach to scoring (the unweighted sum of the number of items correct) functions as an estimate of ability. The relation between ability scale estimates calculated by BILOG and the number correct is shown in Figure 3. These scores were highly correlated ($r = .95$), which suggests that little information would be lost if this score were used as an estimate of ability at the group level. However, nearly a quarter of estimates made using the two methods did not fall in the same quintiles. Disagreement was almost symmetrical, although there was a slight tendency for number correct estimates to underestimate middle level ability and to overestimate the highest level. This result may well simply reflect the coarseness of the distribution of number correct scores. The mean number correct was 51.8 ($SD = 5.8$, range 10–60).

Correlates of Performance on the STW

As the IRT ability estimate and number correct were so highly correlated, the associations with other variables were almost identical for each metric. Accordingly, only results for number correct are reported. STW performance was correlated negligibly with age ($r = .01, p > .05$), although the narrow age range of the sample must be borne in mind. The performance of men did not differ significantly or substantially from that of women, $M_s (SDs) = 52.0 (5.9)$ and $51.6 (5.7)$, respectively; $t(2480) = 1.66, p = .096$.

Table 2 shows that STW performance was moderately related to the cognitive performance and memory tasks included in the study. Years of education, a component of demographic approaches to estimation (Crawford et al., 1989), was most strongly correlated

Table 1
Difficulty (% Correct), Item Response Theory (IRT) Parameters, Item Information, and Fit Statistics for the Spot-the-Word Test

Item no.	Word ^a	% correct	IRT parameter			Loading	Item information		Item fit
			Slope	Threshold	Asymptote		Maximum $I(\theta)_{\max}$	Location $\theta_{\max}$	$\chi^2/df/p$
1	oasis	98.5	1.200	-2.859	0.279	0.77	0.61	-2.69	4.3/4/.36
2	pinnacle	69.1	0.445	0.690	0.500	0.41	0.05	1.33	59.6/9/.00
3	mannerism	93.1	1.544	-1.464	0.366	0.84	0.84	-.31	5.9/5/.32
4	daffodil	99.1	1.185	-3.066	0.405	0.76	0.46	-2.85	3.9/2/.14
5	cyan	49.2	0.692	1.456	0.363	0.57	0.17	1.79	6.1/9/.73
6	sampler	95.7	1.005	-2.188	0.347	0.71	0.37	-1.96	1.3/6/.97
7	necromancy	93.2	0.956	-1.474	0.500	0.69	0.24	-1.18	45.3/6/.00
8	narwhal	40.2	1.686	0.662	0.167	0.86	1.49	0.74	38.0/8/.00
9	monad	84.2	0.457	-1.793	0.297	0.42	0.09	-1.34	7.4/9/.59
10	savage	98.6	0.815	-3.502	0.321	0.63	0.26	-3.24	3.4/5/.64
11	minim	81.8	0.983	-0.386	0.500	0.70	0.25	-0.10	10.5/8/.23
12	mandrake	94.6	0.587	-2.766	0.370	0.51	0.12	-2.36	3.2/7/.87
13	canticle	71.2	1.237	0.306	0.500	0.78	0.40	0.53	10.2/9/.33
14	threnody	31.1	3.815	1.268	0.229	0.97	6.78	1.31	242.2/8/.00
15	banshee	89.4	2.348	-1.105	0.291	0.92	2.27	-1.02	1.4/5/.92
16	shako	49.7	1.295	0.878	0.335	0.79	0.63	1.05	7.7/9/.56
17	paraclete	80.3	0.938	-0.386	0.500	0.68	0.23	-0.08	8.6/8/.37
18	clod	98.2	0.940	-3.004	0.344	0.68	0.33	-2.76	6.7/5/.24
19	rouse	98.8	1.081	-3.023	0.379	0.73	0.40	-2.80	5.5/3/.14
20	goblet	98.5	0.783	-3.434	0.394	0.62	0.20	-3.12	12.7/5/.03
21	viscera	72.3	2.216	-0.260	0.317	0.91	1.92	-0.16	34.1/6/.00
22	almond	99.2	1.039	-3.456	0.333	0.72	0.41	-3.24	2.5/2/.29
23	tarantula	94.9	1.971	-1.475	0.442	0.89	1.16	-1.34	5.3/4/.26
24	rafters	99.0	1.040	-3.211	0.362	0.72	0.39	-2.99	4.4/3/.23
25	archaic	91.0	1.946	-1.158	0.390	0.89	1.27	-1.03	3.3/5/.66
26	obsidian	75.7	1.550	-0.055	0.500	0.84	0.63	0.13	24.2/7/.00
27	zombie	96.3	1.066	-2.277	0.322	0.73	0.44	-2.07	4.5/6/.60
28	pimple	98.4	1.573	-2.340	0.436	0.84	0.75	-2.17	2.7/3/.45
29	static	99.0	1.353	-2.982	0.326	0.80	0.70	-2.82	8.1/2/.02
30	domain	99.0	1.230	-3.050	0.316	0.78	0.59	-2.88	5.6/2/.06
31	livid	97.0	1.592	-1.961	0.411	0.85	0.81	-1.80	1.7/4/.79
32	thrash	99.2	1.286	-3.089	0.371	0.79	0.58	-2.90	5.4/2/.07
33	dross	84.6	2.709	-0.538	0.489	0.94	1.96	-0.43	11.7/5/.04
34	orifice	93.8	2.133	-1.338	0.427	0.91	1.41	-1.22	4.8/4/.31
35	phalanx	88.8	2.635	-0.921	0.403	0.93	2.26	-0.83	12.3/4/.02
36	lapidary	89.9	2.121	-1.082	0.364	0.90	1.60	-0.97	8.3/5/.14
37	palfrey	81.1	0.999	-0.534	0.458	0.71	0.29	-0.26	5.6/8/.69
38	archipelago	93.6	2.673	-1.058	0.500	0.94	1.86	-0.95	14.5/4/.01
39	toga	91.3	1.944	-1.281	0.307	0.89	1.51	-1.17	3.5/5/.62
40	tangible	98.2	1.345	-2.548	0.339	0.80	0.68	-2.38	6.8/3/.08
41	moralist	97.6	1.401	-2.244	0.380	0.81	0.67	-2.07	2.5/4/.65
42	quince	98.1	1.754	-2.211	0.402	0.87	1.01	-2.07	2.0/2/.37
43	epicene	81.1	0.834	-0.459	0.500	0.64	0.18	-0.12	7.2/8/.52
44	gibbon	96.5	0.936	-2.415	0.376	0.68	0.30	-2.16	2.8/5/.73
45	osprey	89.7	1.630	-1.159	0.354	0.85	0.96	-1.02	6.6/6/.36
46	element	99.3	1.310	-3.116	0.384	0.79	0.58	-2.93	6.1/2/.05
47	viridian	58.3	0.870	0.322	0.287	0.66	0.31	0.55	10.4/9/.32
48	onyx	96.0	1.410	-1.867	0.395	0.82	0.66	-1.69	6.4/4/.97
49	plankton	89.8	2.128	-1.068	0.368	0.91	1.59	-0.96	1.5/5/.92
50	akimbo	71.8	1.615	-0.247	0.324	0.85	1.01	-0.11	11.5/8/.18
51	centaur	90.5	2.984	-0.924	0.500	0.95	2.32	-0.83	29.7/4/.00
52	bargain	99.2	1.280	-3.160	0.312	0.79	0.65	-2.99	5.3/2/.07
53	mango	99.1	1.185	3.202	0.326	0.76	0.54	-3.02	4.5/2/.10
54	reticule	84.8	0.897	-0.890	0.442	0.67	0.24	-0.60	8.1/8/.42
55	ontology	92.6	0.366	-3.574	0.340	0.34	0.05	-2.96	37.3/9/.00
56	legerdomain	56.0	3.322	0.549	0.373	0.96	3.85	0.62	52.6/7/.00
57	hoyden	51.5	1.933	0.670	0.331	0.89	1.42	0.78	33.1/8/.00
58	aboriginal	99.6	1.451	-3.336	0.379	0.82	0.72	-3.17	8.0/-/1.00
59	bestiary	75.5	0.585	0.062	0.500	0.50	0.09	0.55	19.1/9/.02
60	albatross	98.5	1.276	-2.732	0.320	0.79	0.63	-2.56	4.1/2/.13

Note. A dash indicates that pattern of responses precluded formal test of goodness of fit.

^a Spot-the-Word items are identified by the real word in each pair.

Figure 1. Observed probability of correct response (circles) compared with the predicted three-parameter item response theory model for four poorly fitting items (r_{IT} is the item-total correlation). Error bars represent the standard errors of the means. Each item is identified by the word from the word-nonword pair.

with the STW ($r = .42$). Digit Span Backward and the SDMT had differentially stronger relationships with the STW than did memory tasks. A hierarchical regression model predicting STW was built, starting with years of education and progressively adding other tests, starting with those most closely identified with general intellectual ability. Sequentially adding Digit Span Backward, the SDMT, delayed recall, and, finally, immediate recall increased the multiple correlation from .42 to .50, .54, .55, and .56. All increments were statistically significant; however, it is apparent that only the first two tests offered meaningful improvements in prediction.

Correlation with the MMSE was .42. Performance on the MMSE as a measure of cognitive status is known to be influenced by educational attainment (Anthony, LeResche, Niaz, Von Korff, & Folstein, 1982); however, controlling for years of education reduced the correlation only slightly ($r = .38$). Performance was unrelated or negligibly related to eyesight, indicators and self-report of physical health, and mental health status (see Table 2).

Discussion

This study is the first analysis of the measurement properties of the STW that is based on a large community sample. IRT

analysis revealed problems with a number of items and with the test. These problems may have escaped detection by those using classical test theory methods. Some items are so easy as to be uninformative. The pseudo word in some items is more attractive to individuals of moderate ability than it is to individuals of lower ability. These problems may well have gone unidentified without the use of IRT. There is a preponderance of items that measure lower levels of ability. Conversely, there is a paucity of reliable items that tap average levels of ability and too few items that measure moderate-to-high ability with precision. The impact of the form of the test information curve would depend on the application. In many situations, the identification of individuals with low premorbid ability is of primary importance, as it is these persons who are most at risk of being wrongly considered to have declined from earlier in life. Thus, this item balance may be considered appropriate for this application. However, the limited ability of the STW to accurately assess higher levels of intellectual ability must be borne in mind: Decline in previously highly functioning individuals may not be detected. This fact could have important implications in clinical applications of the STW, particularly for assessments of subclinical impairment syndromes.

Figure 2. Test information $I(\theta)$ as a function of ability (θ) for a test containing all items and for a test with two poorly fitting items (14, 56) removed.

Problems with individual items notwithstanding, it is clear that the STW discriminates a dimension of cognitive performance. Comprehensive assessment of intellectual ability was precluded by the large-scale epidemiological nature of this survey. However, the

STW was appropriately correlated with the measures of ability on hand. The strength of these associations might be seen as moderate. However, the tests available focused on speeded performance and on memory. In this light, these correlations are in line with

Figure 3. Scatter diagram of estimated ability (θ) by number of correct responses. Gray lines divide quintiles of each measure. Symbol size is proportional to the number of observations represented.

Table 2
Correlations of the Spot-the-Word Test With Demographic Data,
Memory and Performance, and Physical and Mental Health

Variable	r^a
Age	.01
Years of education	.42
Memory and performance	
Digit Span Backward	.37
Symbol Digit Modalities Test	.39
Immediate recall	.31
Delayed recall	.29
Simple reaction time	-.18
Complex reaction time	-.16
Physical health	
Visual acuity	.11
Handgrip	.09
FEV	.15
SF-12 Physical Health Scale	.12
Mental health	
Goldberg Anxiety Scale	-.05
Goldberg Depression Scale	-.06
SF-12 Mental Health Scale	.02

Note. FEV = forced expiratory volume. SF = Short Form.

^aAll correlations except those with age or with SF-12 are statistically significant ($p < .05$). Correlations are based on 2,416–2,482 participants. Variation is due to missing observations.

values that would be expected for an omnibus assessment of ability and are comparable to those observed between the WAIS and tests of this type in older samples (see, e.g., Weschler, 1981). The bivariate and multiple correlations of the STW with education, Digit Span Backward, and symbol–digit substitution suggest the test is substantially aligned with general intellectual ability. The correlation between the STW and the MMSE parallels the findings of moderate differences on the STW in demented and undemented samples (Beardsall & Hupert, 1997; Law & O'Carroll, 1998; McFarlane et al., 2006). This finding implies that caution should be adopted in interpreting performance on the STW as robust to decline. However, all available information regarding robustness comes from cross-sectional comparisons. Risk to dementia is increased by low education and by low premorbid functioning (Valenzuela & Sachdev, 2006). Longitudinal data that demonstrate the stability of the STW in the face of incidence decline in intellectual performance are required to confirm the status of this test as a robust measure of premorbid ability.

Performance on the STW was appropriately unrelated to or only negligibly correlated with age, with eyesight, and with indices of physical health. This information contributes to establishing the absence of measurement artifacts in the STW. However, the relative youth and narrow age range of this older sample must be considered: Reflecting the population at this age, the majority of participants were healthy and well functioning.

Some items, as indicated by the poor fit of the IRT model, are candidates for alteration or removal. At this stage of the development of the STW, we believe, it would be premature to present a revised version of the test based on a subset of the current items. It is certainly the case that the sample is a relatively large one and that IRT parameters are sample independent. However, parameters are not population independent. Our previous IRT modeling of the French adaptation of the NART in different populations (Mackin-

non & Mulligan, 2005; Mackinnon et al., 1999) demonstrated the sensitivity of those items to subtle differences in the groups studied. There is, however, considerable scope for improvement of items and for exploration of alternative formats and content. The multiple choice format is not strictly necessary—respondents could simply be asked to identify actual words in an unpaired list of words and nonwords—and it introduces a number of complications to the test. These include guessing and the possibility of identifying the actual word using a variety of strategies. Recently, promising attempts have been made to adapt the “Spot-the-” format to tap nonlexically based knowledge, which may be robust to decline. These tests have used book titles and country outlines (Scott, de Wit, & Deary, 2006).

Conclusion

Overall, the STW reliably defines a dimension of intellectual ability. Although guessing is inevitable in the binary word identification task, the number correct score is not substantially biased as an indicator of ability. The test correlated appropriately with other measures of ability and was unaffected by participants' visual acuity and physical and mental health. Future refinement of the test could be achieved by eliminating or replacing poorly functioning items. Further, longitudinal research is required to confirm the robustness of the test to incident decline in performance due to dementia and other causes.

References

- Anthony, J. C., LeResche, L., Niaz, U., Von Korff, M. R., & Folstein, M. F. (1982). Limits of the “Mini-Mental State” as a screening test for dementia and delirium among hospital patients. *Psychological Medicine*, 12, 397–408.
- Baddeley, A., Emslie, H., & Nimmo Smith, I. (1992). *The Speed and Capacity of Language Processing Test*. Bury St Edmunds, England: Thames Valley Test.
- Baddeley, A., Emslie, H., & Nimmo Smith, I. (1993). The Spot-the-Word Test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, 32, 55–65.
- Beardsall, L., & Hupert, F. (1997). Short NART, CCRT and Spot-the-Word: Comparisons in older and demented persons. *British Journal of Clinical Psychology*, 36(4), 619–622.
- Christensen, H., Mackinnon, A. J., Jorm, A. F., Henderson, A. S., Scott, R., & Korten, A. E. (1995). ICD-10 mild cognitive disorder: Epidemiological evidence on its validity. *Psychological Medicine*, 25, 105–120.
- Crawford, J. R. (1989). Estimating premorbid intelligence: A review of recent developments. In J. R. Crawford & D. M. Parker (Eds.), *Developments in clinical and experimental neuropsychology* (pp. 55–74). New York: Plenum.
- Crawford, J. R. (1990). Estimating premorbid intelligence by combining the NART and demographic variables. *Personality and Individual Differences*, 11(11), 1153–1157.
- Crawford, J. R., Stewart, L. E., Parker, D. M., Besson, J. A. O., & Cochrane, R. H. B. (1989). Estimation of pre-morbid intelligence: Combining psychometric and demographic approaches. *Personality and Individual Differences*, 10, 793–796.
- Crowell, T. A., Vanderploeg, R. D., Small, B. J., Graves, A. B., & Mortimer, J. A. (2002). Elderly norms for the Spot-the-Word test. *Archives of Clinical Neuropsychology*, 17, 123–130.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test*. San Antonio, TX: Psychological Corporation.
- Dunlap, J. W., De Melo, A., & Cureton, E. E. (1929). The effects of

- different directions and scoring methods on the reliability of a true-false test. *School and Society*, 30, 378–382.
- Folstein, M., Folstein, S., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189–198.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Dordrecht, the Netherlands: Kluwer Nijhoff.
- Hinton-Bayre, A. D., Geffen, G., & MacFarland, K. (1997). Mild head injury and speed of information processing: A prospective study of professional rugby league players. *Journal of Clinical and Experimental Neuropsychology*, 19, 275–289.
- Law, R., & O'Carroll, R. E. (1998). A comparison of three measures of estimating premorbid intellectual level in dementia of the Alzheimer type. *International Journal of Geriatric Psychiatry*, 13(10), 727–730.
- Lehrl, S. (1971). *Mehrfachwahl-Wortschatz-Intelligenztest B (MWT-B)* [Multiple-choice vocabulary intelligence test]. Balingen, Germany: Spitta.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lucas, S. K., Carstairs, J. R., & Shores, E. A. (2003). A comparison of methods to estimate premorbid intelligence in an Australian sample: Data from the Macquarie University Neuropsychological Normative Study (MUNNS). *Australian Psychologist*, 38(3), 227–237.
- Mackinnon, A. J., Christensen, H., Jorm, A. F., Henderson, A. S., Scott, L. R., & Korten, A. E. (1994). A latent trait analysis of an inventory designed to detect symptoms of anxiety and depression in an elderly sample. *Psychological Medicine*, 24, 977–986.
- Mackinnon, A., & Mulligan, R. (2001). The Psychogeriatric Assessment Scales (PAS): Psychometric properties in French and German speaking populations. *International Journal of Geriatric Psychiatry*, 16, 892–899.
- Mackinnon, A., & Mulligan, R. (2005). L'estimation de l'intelligence pré-morbide chez les francophones [Estimation of premorbid intelligence in French speakers]. *L'Encéphale*, 31(1), 31–43.
- Mackinnon, A., Ritchie, K., & Mulligan, R. (1999). The measurement properties of a French language adaptation of the National Adult Reading Test. *International Journal of Methods in Psychiatry*, 8(1), 27–38.
- Marso, D. (2007). *SPSS macro*. Retrieved July 31, 2007, from <http://www.spsstools.net/Syntax/Bootstrap/BootstrapCIforCronbachAlpha.txt>
- McFarlane, J., Welch, J., & Rodgers, J. (2006). Severity of Alzheimer's disease and effect on premorbid measures of intelligence. *British Journal of Clinical Psychology*, 45, 453–463.
- Nelson, H. E. (1982). *National Adult Reading Test (NART) manual*. Berkshire, England: Nfer-Nelson.
- Nelson, H. E., & O'Connell, A. (1982). Dementia: The estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex*, 14, 234–244.
- O'Carroll, R. (1995). The assessment of premorbid ability: A critical review. *Neurocase*, 1, 83–89.
- Paolo, A. M., & Ryan, J. J. (1992). Generalizability of two methods of estimating premorbid intelligence in the elderly. *Archives of Clinical Neuropsychology*, 7–143, 135.
- Psychological Corporation. (2001). *Wechsler Test of Adult Reading*. San Antonio, TX: Author.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.
- Schmand, B. A., Bakker, D., Saan, R. J., & Louman, J. (1991). De Nederlandse Leestest voor Volwassenen: Een maat voor het premorbide intelligentieniveau [The Dutch Adult Reading Test: A measure of premorbid intelligence]. *Tijdschrift voor Gerontologie en Geriatrie*, 22, 15–19.
- Scott, K. M., de Wit, I., & Deary, I. J. (2006). Spotting books and countries: New approaches to estimating and conceptualizing prior intelligence. *Intelligence*, 34(5), 429–436.
- Smith, A. (1973). *Symbol Digit Modalities Test*. Los Angeles: Western Psychological Services.
- Tombaugh, T. N., & McIntyre, N. J. (1992). The Mini-Mental State Examination: A comprehensive review. *Journal of the American Gerontological Society*, 40, 922–935.
- Uttl, B. (2002). North American Adult Reading Test: Age norms, reliability, and validity. *Journal of Clinical and Experimental Neuropsychology*, 24(2), 1123–1137.
- Valenzuela, M. J., & Sachdev, P. (2006). Brain reserve and dementia: A systematic review. *Psychological Medicine*, 36(4), 441–454.
- Ware, J. E., Kosinski, M., & Keller, S. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3), 220–223.
- Watt, K. J., & O'Carroll, R. E. (1999). Evaluating methods for estimating premorbid intellectual ability in closed head injury. *Journal of Neurology, Neurosurgery & Psychiatry*, 66, 474–479.
- Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology*, 19, 87–95.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised*. New York: Psychological Corporation.
- Wechsler, D. (1997). *Wechsler Memory Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.
- Yuspeh, R. L., & Vanderploeg, R. D. (2000). Spot-the-Word: A measure for estimating premorbid intellectual functioning. *Archives of Clinical Neuropsychology*, 15(4), 319–326.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.

Received May 10, 2007

Revision received July 19, 2007

Accepted July 27, 2007 ■