Journal of Clinical and Experimental Neuropsychology
2002, Vol. 24, No. 8, pp. 1123–1137

1380-3395/02/2408-1123$16.00

North American Adult Reading Test: Age Norms, Reliability, and Validity

Bob Uttl
Psychology Department, Oregon State University, Corvallis, OR, USA

ABSTRACT

The North American Adult Reading Test (NAART) is a quickly administered index that is widely used to estimate verbal intellectual ability. We have administered NAART to 351 healthy adults between 18 to 91 years of age to examine psychometric properties of the NAART and to elucidate influence of age, education and gender on NAART performance. The results showed that the NAART is a reliable and valid measure of verbal intelligence, comparable in psychometric properties to the WAIS–R Vocabulary test and with equal psychometric properties in young, middle-aged and older adults. The NAART scores increase across the adult life span (~4.5 points or ~0.5 *SD*) and with education (~1.5 points/year of education) but they are unrelated to gender. The shorter version – the NAART35 – is equally reliable and valid in predicting the WAIS–R Vocabulary. We provide norms as well as various equations for precise predictions of the NAART, the NAART35, and the WAIS–R Vocabulary scores based on age and education. Hierarchical regression analyses demonstrated that the verbal intelligence indexes are useful in predicting and interpreting performance on at least some, but not necessarily all neuropsychological tests, in addition to participants' age.

In clinical practice, the examiner is frequently faced with a problem of estimating an examinee's intellectual functioning prior to the onset of a disease or a traumatic injury leading to intellectual deterioration. The widely used methods for estimating premorbid intellectual functioning include the clinical judgment, the methods based on the demographic information (e.g., Barona index; Barona, Reynolds, & Chastain, 1984), and performance on the current tests of presumed spared abilities such as the WAIS–R Vocabulary subtest (Wechsler, 1981) and various reading tests such as the National Adult Reading Test (NART; see Franzen, Burgess, & Smith-Seemiller, 1997, for a recent review). These methods for estimating premorbid intellectual functioning, including pronunciation-based methods such as the NART, are also widely used in research as quick indexes of participants' intellectual level and as indexes of their semantic memory (Lezak, 1995; Spreen & Strauss, 1991, 1998; Uttl & Graf, 1997).

The National Adult Reading Test was developed and standardized in England (Nelson, 1982). The NART consists of 50 irregularly pronounced English words. An examinee is required to read each word, and the dependent variable is a number of correctly pronounced words. The rationale behind the test is that the reading ability is highly correlated with general intellectual level in the normal population and that in various patient groups, including patients with dementia, reading ability is maintained at or near its premorbid level (Franzen et al., 1997; Nelson, 1982; Spreen & Strauss, 1991, 1998).

Extant research demonstrated that the NART has high internal consistency, generally above .90 (Crawford, Stewart, Garthwaite, Parker, & Besson, 1988; Nelson, 1982), high test-retest reliability of

Address correspondence to: Bob Uttl, Psychology Department, OSU, 204C Moreland Hall, Corvallis, OR 97331-5303, USA. Tel.: +1-541-737-1374. Fax: +1-541-737-3547. E-mail: bob.uttl@orst.edu
Accepted for publication: April 30, 2002.

.98 (Crawford, Parker, Stewart, Besson, & De Lacey, 1989), high interrater reliability of .88 or higher (Crawford et al., 1989), and high validity in estimating WAIS–R Verbal IQ, with correlations ranging from .75 to .85 (Spreen & Strauss, 1998). The NART is less valid in estimating WAIS–R Performance IQ and Full Scale IQ (Spreen & Strauss, 1998). For a recent review of NART see Spreen and Strauss (1998).

The focus of the present investigation is the North American Adult Reading Test (NAART), a modification of the NART developed by Blair and Spreen (1989; see also Spreen & Strauss, 1991) for use with the North American population. Similar to the NART, the NAART consists of 61 different irregular rare words scored for accuracy according to American and Canadian pronunciation rules. Blair and Spreen provided correlations between the NAART scores and WAIS–R Verbal IQ (VIQ), Performance IQ, and Full Scale IQ as well as prediction equations for estimating these IQ indexes from the NAART scores. The internal consistency, measured by Cronbach's α, was high (.94) and a measure of interscorer reliability was almost perfect (.99; Blair & Spreen, 1989). Blair and Spreen's results suggested that the NAART was a successful modification of the NART for the North American population.

In comparison to the NART, however, the NAART has not been extensively researched. Blair and Spreen (1989) developed the NAART on only 66 participants, 18 to 49 years of age. Because of the limited size of their sample, Blair and Spreen's data did not allow us to ascertain magnitude of age, gender and education effects on the NAART performance. Wiens, Bryan, and Crossen (1993) studied the performance of a select group of 302 applicants for civil service positions, primarily women. Only 15 applicants were older than 40 years of age and none were older than 54 years of age. Wiens et al. found no age and gender effects on the NAART scores and only a minimal effect of education (r between years of education and the NAART errors was $-.14$) within the age and education range of their sample. In contrast, Graf and Uttl (1995) and Uttl and Graf (1997) found age-related increase in the NAART accuracy in large samples of 16–84 years old healthy community dwelling adults. Older groups (60–69 and 70–79 years old participants) scored about 0.5 to 0.7 standard deviation higher than younger groups (20–29 and 30–39 years old participants). Graf and Uttl also found that education was positively related to the NAART accuracy. However, because their investigations were not aimed at the NAART specifically, they did not report influence of gender on the NAART performance. In summary, influence of age, education and gender on the NAART scores is still not clear, and at present, there is no large scale adult life-span normative data available for clinicians' and researchers' use that would take account of these factors.

Because of the limited research focused on the NAART, it is also not clear whether psychometric properties – especially reliability and validity – are comparable in young, middle-aged, and older adults. The findings that the NAART scores increase with age raises a possibility that words selected for the NAART could be biased against younger adults. These findings also highlight the need to demonstrate that the NAART predicts other established indexes of verbal intelligence such as the WAIS–R Verbal IQ and/or the WAIS–R Vocabulary scores equally well in young, middle-aged, and older adults.

It has been argued that the NART can be shortened by administering only the first half of the test (i.e., the first 25 items out of 50; Beardsall & Brayne, 1990). Crawford, Parker, Allan, Jack, and Morrison (1991) pointed out that there is a substantial discrepancy between the full length NART scores and the predicted full length NART scores based on the short NART (i.e., the first 25 items). In contrast, when the NART and the short NART were used to predict WAIS–R Verbal IQ, the validity coefficients were comparable (.85 and .83) suggesting that the short NART estimated WAIS–R Verbal IQ equally well. Because shorter tests are generally preferable, it would be useful to find out whether the NAART could also be shortened without a loss of predictive accuracy, especially since the NAART is longer than the NART (61 vs. 50 items).

The goals of the investigation were: (1) to examine influence of age on the NAART scores and to provide age-specific norms based on a large sample of healthy community-dwelling adults

ranging from 18 to 91 years of age, (2) to elucidate influence of gender and education on the NAART scores and, if they predict the NAART performance, to provide prediction equations that include these demographic factors, (3) to provide prediction equations for predicting the WAIS–R Vocabulary scores (both raw and scaled) based on the NAART, (4) to establish reliability and predictive equivalence of the NAART for young, middle-aged, and older adults as well as for the whole sample, and (5) to examine item difficulty and item discrimination indexes to determine if a shorter version of the NAART could be developed with no or only minimal loss of predictive accuracy.

METHOD

Subjects

The subjects were 351 community-living healthy adults, 18–91 years of age, who participated in a study on cognitive aging (see Uttl, Graf, & Cosentino, 2000; Uttl, Graf, & Richter, 2002). They were recruited by means of advertisements in community newspapers, and from a volunteer database at the National Institutes of Health in Bethesda, MD. All participants were tested between May 1997 and October 1998. Participants were paid between $40 and $60, depending on the time required for testing, for attending a single session that lasted from 3 to 4 hr.

The subjects were from the age groups shown in Table 1. Each group spanned 10 years except for the youngest and the oldest, and each group contained about the same number of men and women. The vast majority (98%) of subjects were native English speakers. Six participants who were not native English speakers were bilinguals who spoke English for at least 10 years. Their NAART scores were not different from the NAART scores of the native English speakers, $t(349) = -0.581$, $p = .562$. The table gives demographic and other descriptive data for the entire sample. The participants were comparable in terms of years of formal education, except for the youngest group which averaged significantly fewer years of education than the other groups (by Newman-Keuls, with $\alpha = .05$). The remaining groups were not significantly different from each other.

Table 1 also shows participants' performance on a test of verbal ability – the vocabulary subtest from the revised Wechsler Adult Intelligence Scale – Revised (WAIS–R; Wechsler, 1981). Finally, the table includes performance on the WAIS–R Digit Symbol, a scale used in the computation of the WAIS–R performance IQ index (Wechsler, 1981). Regression analyses showed a significant age-related increase in WAIS–R vocabulary raw scores, consistent with previous findings of a small, positive relation between age and verbal intellectual ability (Graf & Uttl, 1995; Spreen & Strauss, 1991, 1998; Uttl & Graf, 1997). Regression analysis of the Digit Symbol raw scores showed a large and linear age-related decline, consistent with numerous previous findings (Salthouse, 1985, 1988).

Participants also responded to the following question: "How is your overall health at the present time? Excellent, Good, Fair or Poor?" Ratings were translated into values from 1 = poor to 4 = excellent, and these values were averaged across subjects. Table 1 shows that subjects over 60 years of age rated their overall health slightly lower than the younger subjects.

Assessment Instruments

Participants completed a 3- to 4-hr battery of tests focusing on sensation, perception, processing resources, memory, knowledge and language skills. A full description of the entire battery of tests appears in Uttl et al. (2000, in preparation). Included in the battery was the North American Adult Reading test (NAART; Blair & Spreen, 1989; Spreen & Strauss, 1991, 1998) administered according to the instructions published by Spreen and Strauss (1991).

Procedure

Each participant was tested individually, in a single session lasting about 4 hr, in a small office on the main campus of the National Institutes of Health in Bethesda, MD. After giving informed consent, subjects completed a large battery of neuropsychological tests including the NAART.

RESULTS & DISCUSSION

Normative Data and Effects of Age, Education and Gender

Table 2 gives complete normative data for the NAART arranged by midpoint overlapping age groups (Pauker, 1988; Uttl & Graf, 1997; Uttl et al., 2002). Arranging the data in this manner has two main advantages: It maximizes the practical usefulness of the available data and it smoothes out group-specific irregularities in the data set. Tabled values highlight a moderate age-related increase in the NAART accuracy expected from our prior investigations (Graf & Uttl, 1995; Uttl & Graf, 1997) as well as from prior investi-

Table 1. Description of Participants.

		Age Group (years)								Age[a] $F(1,350)$	Age^{2a} $F(1,349)$
		18–19	20–29	30–39	40–49	50–59	60–69	70–79	80–91		
N		13	59	48	53	56	51	48	23		
Females	*n*	8	28	23	27	28	27	24	12		
Males	*n*	5	31	25	26	28	24	24	11		
Occupation											
Retired	*n*			1		6	33	48	23		
Student	*n*	10	35	5			1				
Unemployed	*n*	1	3		2	2	1				
Homemaker	*n*		2	3	4	1					
Semi-skilled	*n*	1	8	8	13	5	3				
Skilled	*n*		6	22	20	33	5				
Professional	*n*		2	3	7	6	4				
Missing data	*n*	1	3	6	7	3	4				
Ethnicity											
Caucasian	*n*	8	42	40	46	52	49	44	23		
African American	*n*	2	2	3	3	2		3			
Asian	*n*	2	10	3	2		1				
Hispanic	*n*	1	2	1	2						
Other + Unknown	*n*		3	1		2	1	1			
Education[1]											
< 12 years	*n*		1				2				
12–15 years	*n*	13	30	20	24	23	24	19	10		
16 years	*n*		21	12	14	15	11	8	7		
> 16 years	*n*		7	16	15	18	14	21	6		
Education[1]	*M*	13.2	15.2	15.6	15.8	16.0	15.2	16.2	15.6	4.58*	5.01*
(# years)	*SD*	0.55	1.54	2.25	2.79	2.65	2.74	2.51	2.63		
Vocabulary[2]	*M*	54.7	55.9	56.0	56.2	58.1	58.0	57.8	58.1	4.33*	0.21
(# correct)	*SD*	10.21	7.34	7.11	8.86	8.25	7.01	6.70	7.60		
Vocabulary[3]	*M*	14.2	12.3	11.7	12.4	13.1	13.2	13.2	13.4		
(scaled scores)	*SD*	3.14	2.36	2.00	2.62	2.55	2.12	1.93	2.25		
Digit Symbol[4]	*M*	70.3	71.8	65.6	60.2	56.6	52.7	46.2	39.7	314.7*	1.03
(#correct)	*SD*	10.48	10.81	10.77	10.42	11.27	9.88	9.17	8.46		
Digit Symbol[5]	*M*	12.8	12.8	12.1	11.8	12.1	12.8	13.8	12.2		
(scaled scores)	*SD*	2.27	2.50	2.53	2.20	2.28	2.36	2.28	2.21		
Overall health rating[6]	*M*	3.38	3.58	3.42	3.43	3.57	3.22	3.13	3.18	17.39*	4.98*
	SD	0.51	0.56	0.58	0.57	0.53	0.55	0.65	0.50		

Note. [a]Age effects were computed by regression analysis.

[1]The number of years attending school/educational institution, whether for a degree or education only; it includes a part-time attendance.

[2]WAIS–R Vocabulary scores (Wechsler, 1981).

[3]WAIS–R Vocabulary scaled scores ($M = 10$, $SD = 3$) calculated using WAIS–R (Wechsler, 1981) age-specific raw-to-scaled scores normative tables. For subjects over 74 years of age, scaled scores were calculated using the last available age group in the manual (70–74 years old).

[4]WAIS–R Digit Symbol scores (Wechsler, 1981).

[5]WAIS–R Digit Symbol scaled scores ($M = 10$, $SD = 3$) calculated using WAIS–R (Wechsler, 1981) age-specific raw-to-scaled scores normative tables. For subjects over 74 years of age, scaled scores were calculated using the last available age group in the manual (70–74 years old).

[6]Overall physical health, self-rated on a 4-point scale: poor = 1, fair = 2, good = 3, excellent = 4.

*$p < .05$.

Table 2. Performance on NAART by Midpoint Overlapping Age Groups, for All Participants and by Years of Education.

		Age group midpoint and range[a]													Age[b]		Age^2[b]	
		20 18–25	25 20–30	30 25–35	35 30–40	40 35–45	45 40–50	50 45–55	55 50–60	60 55–65	65 60–70	70 65–75	75 70–80	80 75–91	r^2	F	Δr^2	F
	n	52	63	55	51	51	59	68	62	59	56	57	52	48				
All Participants																		
	M	38.46	39.90	39.44	38.58	40.05	40.20	41.57	42.88	44.38	42.28	43.15	43.55	43.82	.038	13.93*	.001	0.27
	SD	(9.29)	(8.30)	(8.57)	(9.33)	(10.87)	(10.79)	(9.16)	(8.33)	(8.26)	(9.31)	(9.42)	(8.84)	(8.09)				
By Years of Education																		
≤15 years	M	38.33	40.13	37.02	36.12	37.50	35.14	35.65	38.23	40.11	38.03	37.11	40.08	42.15				
	SD	(9.10)	(7.84)	(9.22)	(9.11)	(10.46)	(9.98)	(8.82)	(8.95)	(9.04)	(9.93)	(10.09)	(10.14)	(8.12)				
	n	36	32	24	21	21	27	29	26	24	29	24	22	22				
≥16 years	M	38.75	39.68	41.32	40.30	41.84	44.46	45.97	46.25	47.31	46.85	47.55	46.10	45.23				
	SD	(9.99)	(8.88)	(7.66)	(9.25)	(10.98)	(9.66)	(6.63)	(5.99)	(6.27)	(5.94)	(5.89)	(6.86)	(7.95)				
	n	16	31	31	30	30	32	39	36	35	27	33	30	26				

Note. [a]The sum of all subjects is greater than 351 due to the manner in which groups were constructed.
[b] Age and age^2 effects were computed by regression analysis.
*$p < .05$.

gations of the closely related NART (Crawford, Nelson, Blackmore, Cochrane, & Allan, 1990; Spreen & Strauss, 1998). A hierarchical regression analysis that used age and age^2 (age^2 was included to capture any possible nonlinear differences across age groups) as predictors revealed moderate linear age-related increase in performance (see Table 2).

To elucidate influence of education and gender, we conducted two hierarchical regression analyses with age, age^2, education and gender as predictors. In order to examine a possibility that a gender and/or education effects may depend on age, we also included interaction terms, Age × Education and Age × Gender. For the first analysis, we allowed education and gender to enter the model on the first step, age and age^2 on the second step, and the interactions on the third step. The results showed that education explained 17.4% of variance, $F(1, 349) = 73.46$, $p < .001$, and age explained an additional 2.2% of variance in NAART scores, $F(1, 348) = 9.66$, $p = .002$. For the second analysis, we allowed age and age^2 to enter the model on the first step, education and gender on the second step, and the interactions on the third step. In this model, age explained 3.8% of the variance, $F(1, 349) = 13.92$, $p < .001$, and education explained an additional 15.8% of variance, $F(1, 348) = 68.33$, $p < .001$. These results indicate that both age and education contribute independently to NAART performance. However, the magnitude of education effect is much larger than the effect of age. In contrast, participant gender did not explain any significant variance in the NAART scores. Since the study had a large statistical power to elucidate even very small effects, it indicates that the effects of gender are either minimal or nonexistent.

Because both education and age have independent effects on NAART performance, Table 2 also includes normative data for participants with 15 or fewer years of education (LoEdu group) and for participants with 16 or more years of education (HiEdu group). In order to optimize the usefulness of our normative data, and to facilitate their application to clinical decision-making, we developed 2 prediction equations. For the first one, we used only age and for the second one we used both age and education as predictors.

$$\text{NAART} = 36.60 + 0.0925 \times \text{Age} \quad (\text{SEE} = 9.15)$$

$$\text{NAART} = 14.07 + 1.518 \times \text{Education} + 0.071 \times \text{Age} \quad (\text{SEE} = 8.38)$$

The prediction equations can be used to calculate the expected score of an individual based on the information available. Corresponding standard error of estimate (SEE) – a value analogous to a standard deviation – can be used to determine how far the individual's obtained score is from the expected/normative score. Assuming we had a person 60 years of age, we can calculate that person's expected NAART score using the first prediction equation. Thus, an expected score is $36.60 + 0.0925 \times 60 = 42.2$ correct. We can also calculate z score equivalent of the person's obtained score or the distance of the obtained score from the expected score in terms of SEE (i.e., standard deviation) by dividing the obtained-minus-expected score by the corresponding SEE. Thus, if our person obtained a score of 61 (a maximum) on the NAART, we would conclude that the person scored $(61 - 42.2)/9.15 = 2.06$ SEEs above the expected score ($z = 2.06$) or in a very superior range. We can also calculate the scaled score standardized to a mean of 10 and standard deviation of 3 by using the following formula: $10 + z \text{ score} \times 3 = 10 + (2.06 \times 3) = 16.2$.

Although our sample does not include sufficient number of individuals from each minority group to develop separate minority specific norms, comparison of age- and education-corrected NAART scores indicate that, relative to the age- and education-specific sample norms, African Americans scored 0.94 *SD* ($n = 15$) below the norms, Asians scored 0.07 *SD* ($n = 18$) below the norms, Hispanics scored 0.66 *SD* ($n = 6$) below the norms, participants from other or unknown ethnic groups scored 0.32 *SD* ($n = 8$) below the norms, and Caucasians scored 0.07 *SD* above the sample norms.

Reliability, Validity and Estimation of WAIS–R Vocabulary Scores

The reliability of the NAART scores, estimated by Cronbach's α, is .93. This value is high, similar to the reliability found by Blair and Spreen (1989) when they first introduced the NAART. The reliability is also comparable to reliabilities found

Table 3. Cumulative Distribution of Differences Between Estimated and Obtained WAIS–R Vocabulary Scores.

Difference ≤	NAART[1]	NAART35[2]
1	0.25	0.26
2	0.38	0.36
3	0.51	0.50
4	0.62	0.62
5	0.74	0.74
6	0.81	0.82
7	0.88	0.88
8	0.91	0.92
9	0.94	0.95
10	0.96	0.95
11	0.97	0.96
12	0.97	0.97
13	0.98	0.98
14	0.99	0.99
15	0.99	0.99

Note. Proportions are based on 351 examinees.
[1]Estimated WAIS–R Vocabulary = 31.30 + 0.622 *NAART.
[2]Estimated WAIS–R Vocabulary = 39.65 + 0.795 *NAART35.

with other established instruments assessing verbal intelligence including the WAIS–R Verbal IQ (Wechsler, 1981), the WAIS–R Vocabulary subtest (Wechsler, 1981), and the NART (Nelson, 1982). Interrater reliability (i.e., correlation between scores obtained from two different scorers) was .93.

To confirm the validity of the NAART as a predictor of the WAIS–R Vocabulary, we have computed correlations between the NAART and the WAIS–R Vocabulary raw scores. The validity coefficient was .75, comparable to previous findings with the NAART (Spreen & Strauss, 1998) as well as with the NART (Crawford et al., 1989). Table 3 shows the cumulative proportion distribution of the differences between estimated and obtained WAIS–R Vocabulary scores.

We have also developed prediction equations for predicting the WAIS–R Vocabulary raw scores and the WAIS–R Vocabulary scaled scores (age-specific) based on the NAART scores. Two equations are provided for each criterion, one based on the NAART only and one based on both the NAART and education. The numbers in parentheses are standard errors of estimate.

$$\text{WAIS–R Vocabulary} = 31.30 + 0.622 \times \text{NAART} \ (\text{SEE} = 5.14)$$

$$\text{WAIS–R Vocabulary} = 25.71 + 0.566 \times \text{NAART} + 0.508 \times \text{Education} \ (\text{SEE} = 5.02)$$

$$\text{WAIS–R Vocabulary Scaled} = 5.383 + 0.179 \times \text{NAART} \ (\text{SEE} = 1.71)$$

$$\text{WAIS–R Vocabulary Scaled} = 4.112 + 0.167 \times \text{NAART} + 0.115 \times \text{Education} \ (\text{SEE} = 1.69)$$

Equivalence of NAART's Psychometric Properties in Young, Middle-Aged and Older Adults

The findings that the NAART scores increase with age raise a possibility that the NAART measures verbal intelligence with a different degree of accuracy in different age groups, that it is differentially valid as a predictor of verbal intelligence, and/or that words selected for the NAART could be biased against younger adults. For this reason, we have divided a sample into three groups – young (age range = 18–39, $n = 120$), middle-aged (age range = 40–59, $n = 109$), and older adults (age range = 60–91, $n = 122$) – and calculated the NAART's psychometric properties for each group separately.

Reliabilties, estimated by Cronbach's α, were comparable: .92 for young, .94 for middle-aged, and .93 for older adults. Validity coefficients – correlations between the NAART and the WAIS–R Vocabulary raw scores – were also comparable: .75 for young, .77 for middle-aged, and .72 for older adults (see Table 4). These findings indicate that NAART measures verbal intelligence with comparable accuracy and validity in the three age groups.

To determine whether or not the NAART is biased, we have calculated a regression line predicting an external criterion – the WAIS–R Vocabulary raw scores – for each age group separately. If the regression lines have comparable slopes and intercepts, we can conclude that the NAART is not biased against any of the three groups (Crocker & Algina, 1986; Kaplan & Saccuzzo, 2001). Table 4 shows intercepts a and slopes b (with standard

Table 4. Results of Regression Analyses Predicting WAIS–R Vocabulary Raw Scores by Age Group.

	Young	Middle	Old	All	All scaled scores
n	120	109	122	351	351
NAART	38.83 (9.07)	41.57 (9.73)	43.23 (8.72)	41.21 (9.32)	
r	.747	.765	.724	.749	.700
r^2	.557	.585	.524	.561	.489
a (intercept)	31.57 (2.04)	29.18 (2.34)	32.67 (2.23)	31.30 (1.25)	5.383 (0.414)
b (slope)	0.623 (0.051)	0.673 (0.055)	0.582 (0.051)	0.622 (.029)	0.179 (0.010)
SEE	5.06	5.55	4.89	5.14	1.71

errors in parentheses) computed for each age group as well as for a complete sample. Neither intercepts nor slopes differed among the groups (largest $t = 1.25$), indicating that the NAART shows no evidence of age bias in predicting the WAIS–R Vocabulary raw scores.

Another way to address an issue of test bias is to examine item difficulties – proportion of examinees passing each item – derived for each age group separately. If the NAART items are unbiased then the item difficulties for young, middle-aged, and older adults ought to correlate very highly with the item difficulties for the other two age groups. Table 5 shows item difficulties for each of the three groups. To identify possibly biased items, we have used the delta measure of item difficulty (Angoff, 1982; Angoff & Ford, 1973; Crocker & Algina, 1986). This method involves calculating deltas for each item and each subgroup, constructing a scatterplot between the deltas and plotting a best fitting straight line to the scatterplot. Next, the absolute value of the distance of each item from the straight line is calculated. If an item deviates sufficiently from the line, it is considered biased. Using this method and a cutoff of 2.0, only 1 item out of 61 – *capon* – showed a possible evidence of bias, with the absolute distance 2.33 from young to middle-aged regression line and 3.57 from young to old regression line.

In combination these results indicate that the NAART is equally accurate and valid in all three age groups and that the NAART measures verbal intelligence free of age-bias. Thus, age-related increase in the NAART scores cannot be attributed to differential validity of the NAART and/or existence of age-related bias against younger groups.

Short Version of NAART (NAART35)

To determine if a shorter version of the NAART could be developed with no or only minimal loss of reliability and predictive accuracy, we have calculated several indexes that allow us to assess (1) how well each item contributes to the reliability and validity of the NAART, and (2) how much different raters agree in classifying each word's pronunciation as correct or incorrect. Table 5 shows item difficulty, corrected item-to-total correlation, discrimination index and Cohen's kappa (Cohen, 1960) for each word. Item difficulty is the proportion of examinees correctly pronouncing each item. Corrected item-to-total correlations measure the degree of relationship between correctly pronouncing each item and the number of all other correctly pronounced items. The discrimination index indicates how well each item discriminates between the low performing 33% of examinees and the top performing 33% of examinees on the NAART. Finally, Cohen's kappa indicates the degree of agreement between two raters following correction for chance agreement; it can range from -1 to $+1$ with -1 indicating perfect disagreement, 0 indicating chance agreement, and 1 indicating perfect agreement. As a guideline for interpreting Cohen's kappa, a value greater than .75 indicates "excellent" agreement, a value between .40 and .75 indicates "fair to good" agreement, and a value less than .40 indicates "poor" agreement (Fleiss, 1971; Kaplan & Saccuzzo, 2001).

To construct the shorter version of the NAART, we have proceeded in a series of steps. First, using Cohen's kappas in Table 5, we have excluded items with less than "excellent" interrater agreement (i.e., items with Cohen's kappa $< .75$). Second, we have excluded items that do not

Table 5. The NAART Item Characteristics.

Item	Word	BS'89[1]	Age group			All	Item-to-total r	Discrimination index d_{U-L}	Inter-scorer reliability κ	Notes
			Young	Middle	Old					
1	debt	0.97	0.96	0.92	0.95	0.94	0.39	0.11	.547	
2	**debris**	0.95	0.93	0.94	0.92	0.93	0.50	0.18	**.902**	
3	aisle	0.95	0.97	0.96	0.98	0.97	0.39	0.08	**.946**	flat ICC
4	reign	0.95	0.97	0.96	0.98	0.97	0.47	0.08	**.946**	
5	depot	0.94	0.94	0.93	0.98	0.95	0.42	0.13	.675	
6	**simile**	0.94	0.91	0.84	0.86	0.87	0.56	**0.34**	**.919**	
7	lingerie	0.92	0.91	0.95	0.98	0.95	0.41	0.16	**.831**	
8	recipe	0.91	0.96	0.95	0.93	0.95	0.40	0.16	**.898**	
9	gouge	0.91	0.93	0.87	0.89	0.90	0.34	0.16	**.750**	
10	heir	0.89	0.90	0.91	0.94	0.92	0.43	0.23	.707	
11	**subtle**	0.89	0.93	0.93	0.91	0.92	0.52	0.23	**.890**	
12	catacomb	0.89	0.98	0.98	0.98	0.98	0.16	0.03	.212	flat ICC
13	**bouquet**	0.88	0.91	0.93	0.96	0.93	0.52	0.18	**.893**	
14	gauge	0.86	0.75	0.93	0.89	0.85	0.26	0.21	**.770**	
15	**colonel**	0.86	0.84	0.93	0.94	0.90	0.43	0.24	**.883**	
16	subpoena	0.86	0.91	0.96	0.95	0.94	0.35	0.10	**.823**	
17	placebo	0.85	0.94	0.93	0.89	0.92	0.34	0.13	**.794**	
18	procreate	0.83	0.94	0.89	0.93	0.92	0.39	0.18	.561	
19	psalm	0.82	0.90	0.94	0.97	0.94	0.38	0.17	**.789**	flat ICC
20	banal	0.80	0.83	0.87	0.84	0.85	0.44	**0.38**	.526	
21	**rarefy**	0.77	0.76	0.81	0.82	0.79	0.56	**0.47**	**.897**	
22	**gist**	0.74	0.88	0.83	0.83	0.85	0.50	**0.40**	**.824**	
23	**corps**	0.71	0.70	0.85	0.88	0.81	0.54	**0.43**	**.879**	
24	**hors d'oeuvre**	0.68	0.68	0.77	0.76	0.74	0.45	**0.44**	**.872**	
25	**sieve**	0.67	0.45	0.61	0.76	0.61	0.49	**0.55**	**.941**	
26	**hiatus**	0.65	0.89	0.84	0.84	0.86	0.54	**0.38**	**.895**	
27	**gauche**	0.65	0.43	0.51	0.45	0.46	0.51	**0.69**	**.868**	
28	**zealot**	0.64	0.72	0.78	0.77	0.75	0.46	**0.48**	**.871**	
29	**paradigm**	0.64	0.79	0.69	0.67	0.72	0.39	**0.39**	**.883**	
30	**façade**	0.61	0.82	0.77	0.84	0.81	0.53	**0.41**	**.896**	
31	**cellist**	0.59	0.65	0.70	0.70	0.68	0.55	**0.60**	**.922**	
32	**indict**	0.59	0.86	0.87	0.93	0.89	0.44	0.24	**.826**	
33	**détente**	0.58	0.38	0.57	0.58	0.51	0.54	**0.72**	**.850**	
34	**impugn**	0.55	0.75	0.78	0.83	0.79	0.60	**0.55**	**.878**	
35	capon	0.53	0.44	0.79	0.91	0.71	0.33	**0.37**	**.807**	age-biased
36	radix	0.52	0.57	0.60	0.62	0.60	−0.02	0.00	**.889**	flat ICC
37	**aeon**	0.52	0.61	0.68	0.64	0.64	0.36	**0.50**	**.758**	
38	**epitome**	0.50	0.68	0.70	0.65	0.67	0.43	**0.47**	**.799**	
39	equivocal	0.48	0.85	0.90	0.93	0.89	0.47	0.25	.700	
40	**reify**	0.44	0.49	0.61	0.66	0.59	0.48	**0.66**	**.798**	
41	**indices**	0.42	0.55	0.61	0.66	0.61	0.55	**0.65**	**.920**	
42	**assignate**	0.41	0.35	0.50	0.61	0.49	0.39	**0.48**	**.786**	
43	**topiary**	0.41	0.60	0.68	0.66	0.64	0.52	**0.57**	**.863**	
44	**caveat**	0.39	0.63	0.73	0.78	0.71	0.59	**0.67**	**.809**	
45	superfluous	0.36	0.60	0.60	0.68	0.63	0.53	**0.68**	.713	
46	**leviathan**	0.35	0.60	0.45	0.55	0.54	0.39	**0.54**	**.810**	
47	prelate	0.27	0.80	0.83	0.88	0.84	0.20	0.19	.546	flat ICC

Table 5. (continued).

Item	Word	BS'89[1]	Age group			All	Item-to-total r	Discrimination index $d_{U\text{-}L}$	Inter-scorer reliability κ	Notes
			Young	Middle	Old					
48	**quadruped**	0.26	0.41	0.45	0.52	0.46	0.49	**0.64**	**.843**	
49	**sidereal**	0.26	0.30	0.38	0.39	0.35	0.45	**0.61**	**.844**	
50	**abstemious**	0.20	0.30	0.29	0.40	0.33	0.41	**0.52**	**.840**	
51	**beatify**	0.17	0.28	0.51	0.50	0.43	0.50	**0.66**	**.930**	
52	**gaoled**	0.15	0.11	0.23	0.36	0.23	0.42	**0.48**	**.886**	
53	**demesne**	0.15	0.09	0.25	0.33	0.22	0.42	**0.45**	**.828**	
54	**syncope**	0.15	0.22	0.34	0.18	0.24	0.29	**0.30**	**.834**	
55	**ennui**	0.14	0.23	0.20	0.36	0.26	0.44	**0.56**	**.890**	
56	drachm	0.11	0.08	0.11	0.12	0.11	0.22	0.21	.678	flat ICC
57	cidevant	0.11	0.07	0.06	0.08	0.07	0.21	0.13	.289	flat ICC
58	epergne	0.09	0.24	0.17	0.25	0.22	0.33	**0.36**	.521	
59	vivace	0.08	0.23	0.22	0.30	0.25	0.40	**0.48**	.536	
60	talipes	0.08	0.14	0.19	0.28	0.21	0.15	0.21	**.757**	
61	synecdoche	0.06	0.11	0.07	0.05	0.08	0.24	0.16	**.893**	
	Cronbach's α	.92	.92	.94	.93	.93				

Note. Words printed in bold constitute a short version of NAART called NAART35. Bold print also highlights item discrimination indexes $\geq .30$ and Cohen's kappas $\geq .75$.
[1]Values reported by Blair and Spreen (1989).

discriminate well between examinees scoring low versus high on the NAART; we have excluded (a) items with flat item characteristic curves and (b) items with low discrimination index ($< .30$) unless their item characteristic curves showed that they discriminated well for either low or high scoring examinees. Third, we have excluded "capon" because as we mentioned above, this item may be age-biased. The resulting shorter version of the NAART includes 35 out of 61 original items identified by bold print in Table 5. The items included in the short version of the NAART (NAART35) have excellent interrater agreement, generally high discrimination indexes and item-to-total correlations, and their item characteristics curves show that these items provide information about individual differences.

Table 6 gives complete normative data for the NAART for the whole sample as well as normative data for participants with 15 or fewer years of education (labeled LoEdu group) and for participants with 16 or more years of education (labeled HiEdu group). In order to optimize the usefulness of our normative data, and to facilitate their application to clinical decision-making, we developed 2 prediction equations. For the first one, we used only age and for the second one we used both age and education as predictors.

$$\text{NAART35} = 19.27 + 0.065 \times \text{Age} \ (\text{SEE} = 7.08)$$

$$\text{NAART35} = 1.76 + 1.180 \times \text{Education} + 0.047 \times \text{Age} \ (\text{SEE} = 6.47)$$

The reliability of the NAART35 scores, estimated by Cronbach's α, is .92. The reliability for young, middle-aged, and older adults is .90, .93, and .92, respectively. These values are comparable to the full length NAART. The correlation between the NAART35 and the NAART was .98. To confirm a validity of the NAART35 as a predictor of the WAIS–R Vocabulary, we have computed correlations between the NAART35 and the WAIS–R Vocabulary raw scores. The validity coefficient was .76, a value indistinguishable from the validity of the full length NAART. Table 3 shows the cumulative proportion distribution of the differences between estimated and obtained WAIS–R Vocabulary scores based on the short version of the NAART. The cumulative distributions show that estimates based on the NAART35 are at least as good as the estimates based on the full version of the NAART.

Table 6. Performance on NAART35 by Midpoint Overlapping Age Groups, for All Participants and by Years of Education.

		Age group midpoint and range[a]													Age[b]		Age^{2b}	
		20 18–25	25 20–30	30 25–35	35 30–40	40 35–45	45 40–50	50 45–55	55 50–60	60 55–65	65 60–70	70 65–75	75 70–80	80 75–91	r^2	F	Δr^2	F
	n	52	63	55	51	51	59	68	62	59	56	57	52	48				
All Participants																		
	M	20.60	21.76	21.39	20.59	21.80	21.72	22.75	23.76	24.84	22.90	23.70	24.18	24.51	.032	11.36*	<.001	0.08
	SD	(6.63)	(6.04)	(6.32)	(7.11)	(8.48)	(8.29)	(7.27)	(6.66)	(6.32)	(7.31)	(7.59)	(7.08)	(6.59)				
By Years of Education																		
≤15 years	*M*	20.59	22.00	19.54	18.71	19.74	17.82	17.99	20.05	21.56	19.51	18.87	21.41	23.25				
	SD	(6.52)	(5.54)	(6.53)	(6.82)	(8.19)	(7.73)	(6.87)	(7.03)	(6.81)	(7.81)	(8.27)	(7.96)	(6.72)				
	n	36	32	24	21	21	27	29	26	24	29	24	22	22				
≥16 years	*M*	20.63	21.52	22.82	21.91	23.23	25.01	26.28	26.44	27.09	26.41	27.06	26.21	25.58				
	SD	(7.10)	(6.61)	(5.87)	(7.13)	(8.52)	(7.37)	(5.34)	(4.94)	(4.91)	(4.75)	(4.85)	(5.68)	(6.41)				
	n	16	31	31	30	30	32	39	36	35	27	33	30	26				

Note. [a]The sum of all subjects is greater than 351 due to the manner in which groups were constructed.
[b]Age and age^2 effects were computed by regression analysis.
*$p < .05$.

The prediction equations for predicting the WAIS–R Vocabulary raw scores and the WAIS–R Vocabulary scaled scores (age-specific), based on the NAART35 scores are provided for each criterion, one based on the NAART35 only and one based on both the NAART35 and Education.

$$\text{WAIS–R Vocabulary} = 38.67 + 0.811 \times \text{NAART35} (\text{SEE} = 5.11)$$

$$\text{WAIS–R Vocabulary} = 32.50 + 0.740 \times \text{NAART35} + 0.500 \times \text{Education} (\text{SEE} = 5.00)$$

$$\text{WAIS–R Vocabulary Scaled} = 7.52 + 0.233 \times \text{NAART35} (\text{SEE} = 1.71)$$

$$\text{WAIS–R Vocabulary Scaled} = 6.12 + 0.217 \times \text{NAART35} + 0.114 \times \text{Education} (\text{SEE} = 1.69)$$

Finally, we also provide a prediction equation for predicting the NAART from the NAART35.

$$\text{NAART} = 12.39 + 1.282 \times \text{NAART35} (\text{SEE} = 1.54)$$

Convergent and Predictive Validity of WAIS–R Vocabulary, NAART, and NAART35

Table 7 shows correlations between measures of verbal intelligence – the WAIS–R Vocabulary, the NAART and the NAART35 – and demographic variables (age, education, and gender) and measures of various cognitive abilities including the WAIS–R Digit Symbol (Wechsler, 1981), the Trail Making Test (Lezak, 1995), the Rey Auditory Verbal Learning Test Trials 1 to 3 (Spreen & Strauss, 1998), and the Paired Associate Test Trials 1 and 2 (Uttl et al., 2002).

The pattern of correlations indicates that the NAART and the NAART35 have identical or almost identical correlations with both demographic variables and measures of cognitive abilities (the largest absolute difference in correlations is less then 0.03). The pattern of correlations also indicates that convergent validities of the NAART/NAART35 and WAIS–R Vocabulary are comparable for demographic variables, but that the WAIS–R Vocabulary validities are slightly higher, relative to the NAART/NAART35 validities, for measures of complex attention and episodic explicit memory (on the average, validities are 0.09 higher for the WAIS–R Vocabulary than for the NAART/NAART35). This small difference in validity coefficients is expected because the WAIS–R Vocabulary test requires examinees to define the meaning of the words, and therefore, their scores on the WAIS–R Vocabulary depend more on attention and higher cognitive processes than pronunciation-based measures of verbal intelligence such as the NAART/NAART35.

Small to moderate convergent validity coefficients in Table 7 may suggest that the indexes of verbal intelligence are not useful in predicting and interpreting performance on various neuropsychological tests, including tests of attention and episodic memory in Table 7. However, such conclusion is not warranted. Previous research has shown that performance on measures of crystallized intelligence such as the NAART and the WAIS–R Vocabulary increases across the adult life span whereas performance on measures of fluid intelligence such as the WAIS–R Digit Symbol and the Trail Making Test declines steeply with age, thereby attenuating correlations between the indexes of verbal intelligence and performance on these neuropsychological tests. To assess usefulness of the verbal intelligence measures, Table 7 also shows a proportion of variance explained in each neuropsychological test by age only (r^2_{age}) and an increase in the proportion of variance explained by each of the three indexes of verbal intelligence after age was taken into account (Δr^2). The Δr^2 shown in Table 7 indicate that the indexes of verbal intelligence are useful in estimating performance on at least the tests of higher cognitive functions such as episodic explicit memory. To illustrate, whereas age explains 7 and 14% of variance on Trial 1 and Trial 2 of the verbal paired associate test, the NAART explains additional 11 and 14% on Trial 1 and Trial 2, respectively.

CONCLUDING COMMENTS

The NAART is a reliable and valid measure of verbal intelligence, comparable in psychometric properties to WAIS–R Vocabulary test. The

Table 7. Convergent and Predictive Validity of WAIS-R Vocabulary, NAART, and NAART35.

	WAIS–R[8] Vocabulary	NAART[9]	NAART35	r^2_{age}[10]	Δr^2* WAIS–R[8] Vocabulary	Δr^2* NAART	Δr^2* NAART35
Age	**0.11**	**0.20**	**0.18**				
Education years[1]	**0.45**	**0.42**	**0.42**				
Gender/Female[2]	−0.02	0.05	0.04				
WAIS-R Digit Symbol[3]	**0.15**	0.05	0.05	**.47**	**.05**	**.03**	**.03**
Trail Making Test Form A[4]	−0.04	−0.01	−0.02	**.37**	**.01**	**.02**	**.01**
Trail Making Test Form B[5]	**−0.15**	−0.09	−0.09	**.37**	**.05**	**.04**	**.04**
Rey VLT Trial A1[6]	**0.28**	**0.17**	**0.17**	**.23**	**.11**	**.07**	**.07**
Rey VLT Trial A2[6]	**0.30**	**0.22**	**0.22**	**.26**	**.10**	**.09**	**.09**
Rey VLT Trial A3[6]	**0.29**	**0.19**	**0.19**	**.25**	**.11**	**.07**	**.07**
Paired Associate Trial 1[7]	**0.29**	**0.23**	**0.23**	**.07**	**.11**	**.08**	**.08**
Paired Associate Trial 2[7]	**0.34**	**0.26**	**0.27**	**.14**	**.14**	**.11**	**.11**

Note. Correlations and Δr^2 printed in bold are significant at $p < .05$.

[1]A number of years attending school including part time attendance (self-reported).

[2]Gender was coded as Female = 1 and Male = 0.

[3]WAIS–R Digit Symbol in ms/item (Wechsler, 1981).

[4]Trail Making Test Part A.

[5]Trail Making Test Part B.

[6]Rey Verbal Learning Test, Trials A1 to A3 (Spreen & Strauss, 1998).

[7]Paired Associate Test (Uttl, Graf, & Richter, 2002).

[8]Number correct on WAIS–R Vocabulary subtest (Wechsler, 1981).

[9]Number correct on NAART (Spreen & Strauss, 1991, 1998).

[10]Proportion variance explained in a neuropsychological test by age only.

*Increase in the proportion variance explained in a neuropsychological test after taking into account effects of age.

NAART is equally reliable and valid for young, middle-aged and older adults. The NAART scores show a moderate increase across the adult life span (~4.5 points or ~0.5 *SD*) and with education (~1.5 points for each year of education). In contrast, the NAART scores show no clinically meaningful effects of gender.

The NAART can be shortened to 35 items with no loss of psychometric properties. The shorter version, called the NAART35, is equally reliable and valid in predicting the WAIS–R Vocabulary scores relative to the full length of the NAART. The feature advantage of the NAART35 is that it is shorter to administer, easier to score, and less demanding on a scorer's time. The NAART35, however, still needs to be cross-validated in an independent sample.

All three measures of verbal intelligence – the WAIS–R Vocabulary, the NAART and the NAART35 – show comparable relationships to demographic variables (age, gender, and education) and to measures of various cognitive functions including attention, complex attention, and explicit episodic memory.

The main limitation of our study is that participants were volunteers rather than randomly selected adults from U.S. population. However, this limitation is shared with all other studies involving human subjects, including normative studies such as the widely used Wechsler Adult Intelligence Scales (Wechsler, 1981, 1997) and the Wechsler Memory Scales (Wechsler, 1987, 1997). To assess possible negative impacts of nonrandom selection procedure, we have collected background data on our participants, including their occupation, ethnicity, education and widely used indexes of verbal intelligence. These data suggest that our sample is comparable to general population.

The WAIS–R Vocabulary and Digit Symbol scaled scores shown in Table 2 may raise concerns about the generalizability of the findings reported in this article; they indicate that our sample scored

about 0.66 *SD* above the WAIS–R normative sample (Wechsler, 1981), suggesting that our sample is special, higher functioning than the general population. However, this conclusion does not seem warranted. The strongest argument against it is the mounting evidence that performance on various measures of intelligence, including the WAIS–R, is rising at a rate of 3 to 4 IQ points per decade. This has been highlighted in several prominent studies and reviews (see Flynn, 1984, 1987, 1999; Fuggle, Tokar, Grant & Smith, 1993; Lynn & Pagliari, 1994; Matarazzo, 1990; Neisser, 1998; Uttl & Van Alstine, 2000). To illustrate, the sample used in 1985/1986 for norming the WMS–R averaged 103.9 on the WAIS–R (Wechsler, 1987), compared to the WAIS–R average of 100 established for the normative sample tested between 1976 and 1980 (Wechsler, 1981). On the California Verbal Learning Test, a recent normative sample produced a WAIS–R full scale IQ estimate of 116.3 for healthy elderly persons (Paolo, Trösner, & Ryan, 1997). Tested prior to 1992, a MAYO clinic normative sample produced estimates of 105.5 for the WAIS–R Verbal IQ and 107.5 for the WAIS–R Performance IQ (Ivnik et al., 1992). In line with the score inflation revealed by these investigations, if our sample was similar to existing normative samples, it should show WAIS–R Verbal IQ scores between 106 and 108 by a conservative estimate, or IQs of about 111 by a liberal estimate. The WAIS–R Vocabulary scaled scores in Table 2 are clearly consistent with these estimates obtained from preexisting normative samples. It appears that our sample did not benefit from superior cognitive skills, and therefore, the findings reported in this article can be generalized to the general population.

ACKNOWLEDGMENTS

This research was supported by the Henry M. Jackson Foundation's funding of Bob Uttl, by an operating grant from the Natural Sciences and Engineering Research Council of Canada to P. Graf, and by equipment loans from Alfalab Research Inc. Part of this research was conducted while Bob Uttl was at the National Institutes of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD.

I thank Joy Bonerba, Stephanie Cosentino, Elizabeth Daniels, Victoria Pharr, Pilar Santacruz, Kristin Stover, Carolyn Pilkenton-Taylor and Cory Van Alstine for assisting with the project.

REFERENCES

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting item bias*. Baltimore: Johns Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973). Item race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–105.

Barona, A., Reynolds, C.R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS–R. *Journal of Consulting and Clinical Psychology, 52*, 885–887.

Beardsall, L., & Brayne, C. (1990). Estimation of verbal intelligence in an elderly community: A prediction analysis using a shortened NART. *British Journal of Clinical Psychology, 29*, 83–90.

Blair, J.R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3*, 129–136.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Crawford, J.R., Nelson, H.E., Blackmore, L., Cochrane, R.H.B., & Allan, K.M. (1990). Estimating premorbid intelligence by combining the NART and demographic variables: An examination of the NART standardization sample and supplementary equations. *Personality and Individual Differences, 11*, 1153–1157.

Crawford, J.R., Parker, D.M., Allan, K.M., Jack, A.M., & Morrison, F.M. (1991). The short NART: Cross-validation, relationship to IQ and some practical considerations. *British Journal of Clinical Psychology, 30*, 223–229.

Crawford, J.R., Parker, D.M., Stewart, L.E., Besson, J.A.O., & De Lacey, G. (1989). Prediction of WAIS/IQ with the National Adult Reading Test: Cross-validation and extension. *British Journal of Clinical Psychology, 28*, 267–273.

Crawford, J.R., Stewart, L.E., Garthwaite, P.H., Parker, D.M., & Besson, J.A.O. (1988). The relationship between demographic variables and NART performance in normal subjects. *British Journal of Clinical Psychology, 27*, 181–182.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Hartcourt Brace Jovanovich.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382.

Flynn, J.R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29–51.

Flynn, J.R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Flynn, J.R. (1999). Searching for justice. The discovery of IQ gains over time. *American Psychologist, 54*, 5–20.

Franzen, M.D., Burgess, E.J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid functioning. *Archives of Clinical Neuropsychology, 12*, 711–738.

Fuggle, P.W., Tokar, S., Grant, D.B., & Smith, I. (1993). Rising IQ scores in British children: Recent evidence. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 33*, 1241–1247.

Graf, P., & Uttl, B. (1995). Component processes of memory: Changes across the adult lifespan. *Swiss Journal of Psychology, 54*, 113–130.

Ivnik, R.J., Malec, J.F., Smith, G.E., Tangalos, E.G., Petersen, R.C., Kokmen, E., & Kurkland, L.T. (1992). Mayo's older Americans normative studies. WAIS–R norms for ages 56 to 97. *Clinical Neuropsychologist, 6*, 1–30.

Kaplan, R.M., & Saccuzzo, D.P. (2001). *Psychological testing* (5th ed.). Belmont, CA: Wadsworth.

Lezak, M. (1995). *Neuropsychological assessment (3rd ed.)*. New York, Oxford: Oxford University Press.

Lynn, R., & Pagliari, C. (1994). The intelligence of American children is still rising. *Journal of Biosocial Science, 26*, 65–67.

Matarazzo, J.D. (1990). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Nelson, H.E. (1982). National Adult Reading Test (NART). Windsor, Berkshire, England: The NFER-NELSON Publishing Company.

Paolo, A.M., Trösner, A.I., & Ryan, J.J. (1997). California Verbal Learning Test: Normative data for the elderly. *Journal of Clinical and Experimental Neuropsychology, 19*, 220–234.

Pauker, J.D. (1988). Constructing overlapping cell tables to maximize the clinical usefulness of normative test data: Rationale and an example from neuropsychology. *Journal of Clinical Psychology, 44*, 11–17.

Salthouse, T.A. (1985). *Theory of cognitive aging*. Amsterdam: North-Holland.

Salthouse, T.A. (1988). Resource-reduction interpretation of cognitive aging. *Developmental Reviews, 8*, 238–272.

Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests*. New York, NY: Oxford University Press.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests*. New York, NY: Oxford University Press.

Uttl, B., & Graf, P. (1997). Color-word Stroop test performance across the adult life-span. *Journal of Clinical and Experimental Neuropsychology, 19*, 405–420.

Uttl, B., Graf, P., Bonerba, J., Santacruz, P., Stover, K., Cosentino, S., & Pharr, V. (in preparation). *Elementary components of speed and executive functions mediate age-related changes in episodic verbal memory*.

Uttl, B., Graf, P., & Cosentino, S. (2000). Exacting assessments: Do older adults fatigue more quickly? *Journal of Clinical and Experimental Neuropsychology, 22*, 496–507.

Uttl, B., Graf, P., & Richter, L.K. (2002). Verbal Paired Associates Test. *Archives of Clinical Neuropsychology, 17*, 569–583.

Uttl, B., & Van Alstine, C. (2000, November). In 2050, a typical volunteer in psychology experiments will be a genius. Orlando, FL: National Academy of Neuropsychology.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale – Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale – Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *WAIS-III Administration and Scoring Manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale – III*. San Antonio, TX: Psychological Corporation.

Wiens, A.N., Bryan, J.E., & Crossen, J.R. (1993). Estimating WAIS–R FSIQ from the National Adult Reading Test – Revised in normal subjects. *The Clinical Neuropsychologist, 7*, 70–84.