Ann. of Dyslexia (2009) 59:133–149
DOI 10.1007/s11881-009-0027-8

Measuring adult literacy students' reading skills using the Gray Oral Reading Test

Daphne Greenberg · Hye Kyeong Pae · Robin D. Morris · Mary Beth Calhoon · Alice O. Nanda

Received: 16 July 2008 / Accepted: 12 May 2009 / Published online: 21 July 2009

Abstract There are not enough reading tests standardized on adults who have very low literacy skills, and therefore tests standardized on children are frequently administered. This study addressed the complexities and problems of using a test normed on children to measure the reading comprehension skills of 193 adults who read at approximately third through fifth grade reading grade equivalency levels. Findings are reported from an analysis of the administration of Form A of the Gray Oral Reading Tests—Fourth Edition (Wiederholt & Bryant, 2001a, b). Results indicated that educators and researchers should be very cautious when interpreting test results of adults who have difficulty reading when children's norm-referenced tests are administered.

Keywords Adult literacy assessment · GORT · Test result interpretation

Fourteen percent of the adult population has difficulty with most adult printed materials, while close to another 29% of the adult population has difficulty with any reading task that is not very basic (Kutner, Greenberg, Jin, Boyle, Hsu, & Dunleavy, 2007). Therefore, a

D. Greenberg (✉) · M. B. Calhoon
Department of Educational Psychology and Special Education, Georgia State University, P.O. Box 3979, Atlanta, GA 30302-3979, USA
e-mail: dgreenberg@gsu.edu

M. B. Calhoon
e-mail: mbcalhoon@gsu.edu

H. K. Pae
Division of Teacher Education, University of Cincinnati, P.O. Box 210022, Cincinnati, OH 45221-0022, USA
e-mail: hye.pae@uc.edu

R. D. Morris
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302-5010, USA
e-mail: robinmorris@gsu.edu

A. O. Nanda
68 Brandon Ridge Dr., Sandy Springs, GA 30328, USA
e-mail: alicenanda@gmail.com

significant number of adults in the United States have difficulty with materials encountered in their houses, neighborhoods, and workplaces. A significant number of these adults may have learning disabilities. For example, prior research conducted by Greenberg, Ehri, and Perin (1997, 2002) indicate that adults who attend adult literacy programs often have problems with phonological awareness and decoding that are typical of individuals with reading disabilities. In spite of the high percentage of adults with low reading skills and the suggested prevalence of learning disabilities within this population (Patterson, 2008), a paucity of research exists on this population (Venezky & Sabatini, 2002).

The impetus for this study came from the context of a larger study analyzing the effectiveness of different reading instructional approaches for adult literacy learners[1]. One of the difficulties in conducting adult literacy research is the lack of commercially standardized tests normed on adults with very low literacy skills. Most tests used in research with adult literacy students are often tests standardized on children and therefore associated with typical childhood developmental units such as reading grade equivalency (RGE) levels (Sabatini, Venezky, Kharik, & Jain, 2000). Researchers are critical of using RGE levels to characterize adult reading because they were developed on children's performance (Perin, 1991) and are typically based on material taken from children's basal readers. Therefore, in addition to the typical measurement limitations of RGE levels cited by test theorists (e.g., Anastasi & Urbina, 1997; Crocker & Algina, 1986), additional problems exist in the scaling and interpretation of RGE levels with adults. Due to the paucity of appropriate tests for adults who struggle with reading, researchers often administer tests standardized on children (Sabatini et al., 2000). This study reports on difficulties that may occur when following this practice.

Information regarding test characteristics of special needs populations is critically needed (Costenbader & Adams, 1991). Research on test characteristics of adult literacy students will not only help researchers but will help all persons charged with assessing the reading skills of adults. For example, a survey conducted by Stevens and Price (1999) found that approximately 78% of neuropsychologists administered reading measures as part of their neuropsychological examinations of adults. Close to 12% reported using the Gray Oral Reading Tests (GORT), the test that is the focus of this study.

This article reports results from a study of adults who identify words on the third through fifth grade RGE levels on the Woodcock Johnson III Letter-Word Identification Test of Achievement (Woodcock, McGrew, & Mather, 2001). Specifically, findings are reported from analysis of the administration of Form A of the Gray Oral Reading Tests—Fourth Edition (GORT-4; Wiederholt & Bryant, 2001a). There are no known research studies that have analyzed the administration of the GORT-4 on adult literacy students who identify words between the third and fifth grade levels.

Adults reading at elementary school grade levels often have bad memories of educational settings as well as of testing, particularly having to read aloud, and therefore often consider testing as an anxiety-provoking situation (D'Amico-Samuels, 1991). In order to increase the adult learner's comfort level and success rate, administrations of tests are often started at a lower level than is typically recommended. This philosophy is in accordance to suggestions in the GORT-4 manual, which specifies that for those with poor reading ability, a lower entry point should be initiated. Therefore, because of the nature of this study's sample, all participants started with story 1 instead of starting with story 3

[1] Research supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institute for Literacy, and the U.S. Department of Education - grant# R01 HD43801-01.

(recommended entry level for third and fourth RGE levels) or story 5 (recommended entry level for fifth RGE levels).

The purpose of this study is to evaluate the utility of the child and adolescent normed GORT-4 for adults with low literacy proficiency. This study reports on the performance of adult literacy students on the GORT-4. There are two aspects to this study. The first aspect addresses the general question: When starting with story 1, is the GORT a psychometrically appropriate test for adults who recognize words at the third to fifth grade levels? Answers to this question are important when using a test normed on children with such an adult sample.

The second aspect of this study relates to a finding that was uncovered very early on in the research. In the early stages of test administration, many of the participants in this study established a ceiling at story 1 and/or story 2. This was surprising given the fact that stories 1 and 2 are the recommended entry points for students in grades one and two and this study's participants were all identifying words between the third and fifth grade levels. We therefore decided to continue to administer the passages in order until the next ceiling was established. Much to our surprise, we found that many of the participants who established a ceiling at story 1 and/or story 2 established a basal at or after story 3. This study reports on the performance of these participants. Specifically, what happens when results are analyzed for those participants who reach a ceiling at story 1 and/or story 2, but establish a basal at or after story 3?

Method

Participants

This study was part of a larger project investigating the effectiveness of reading interventions for adult literacy students[2]. Part of the larger project included administration of a pretest battery of reading and reading-related assessments. The participants were 193 adults who scored between the third and fifth grade levels on the Letter-Word Identification subtest of the Woodcock Johnson III Tests of Achievement (Woodcock et al., 2001). The participants attended adult literacy programs in a southeastern city and volunteered to be part of this research study. As indicated by the "Overall" column in Table 1, the majority of the participants were African American, native speakers of English, female, with a mean age of 30.70 years (ranging from 16 years 6 months to 72 years 9 months), and possessing an average educational attainment level of tenth grade. Fifteen percent of the sample was employed full-time, with an additional 13% working part-time. Thirty-four percent of the sample reported that they received some form of governmental benefit (such as supplemental security income, food stamps, and WIC supplemental nutrition benefits).

The GORT-4 is normed on individuals up to the age of 18 years, 11 months. The study's sample included 39 participants between the ages of 16 years and 18 years 10 months, and therefore, preliminary analyses investigated whether differences existed between this age group and those older. Results indicated that there were no significant differences between the participants who were below 19 years of age and those who were 19 or older in their rate, accuracy, fluency, or comprehension GORT-4 scores. Therefore, results were not differentiated by age in the analyses. In addition, 63 speakers of English as a second language (ESL) were included in the sample. There were no significant differences between the ESL and native English-speaking participants in terms of their demographic characteristics nor in

[2] See footnote 1.

Table 1 Demographic characteristics for the overall group, the atypical pattern group, and the non-atypical pattern group

	Overall (n=193)		Atypical (n=97)		Non-atypical (n=96)	
Characteristics	Freq.	%	Freq.	%	Freq.	%
Race (ns)						
African American	140	72.5	72	74.2	68	70.8
Hispanic	20	10.4	7	7.2	13	13.5
Caucasian	17	8.8	10	10.3	7	7.3
Asian	13	6.7	6	6.2	7	7.3
Other/mixed	3	1.6	2	2.1	1	1.0
First Language (ns)						
English	130	67.4	67	69.1	63	65.6
Other	63	32.6	30	30.9	33	34.4
Gender (ns)						
Female	122	63.2	58	59.8	64	66.7
Male	71	36.8	39	40.2	32	33.3
	Mean	SD	Mean	SD	Mean	SD
Age (ns)	30.70	13.71	31.24	14.13	30.15	13.32
Education level (ns)	10.70	2.40	10.46	2.37	10.94	2.42

ns the atypical and non-atypical pattern groups did not differ significantly

terms of their GORT-4 Rate, Fluency, and Comprehension performances. However, as noted in Table 2, a significant difference was observed in Accuracy performance with the native English-speaking participants performing better than the ESL participants. Therefore, results will be presented separately for both ESL and native English-speaking participants.

Procedure

Trained graduate research assistants administered the test battery. Prior to testing the adult literacy participants, the project's psychometrist extensively trained the graduate research assistants. The training included sensitivity to testing adults who have difficulty reading, as well as very specific instructions on how to administer each assessment protocol for this population. Prior to solo test administration, each research assistant practiced with the psychometrist, observed the psychometrist administer the tests, and then was observed by the psychometrist while administering the battery of tests to an adult literacy student. The psychometrist scored all tests, and another scorer independently verified her scores. Finally, data were independently double-entered and compared to catch any data entry errors. Data analysis is based on the raw scores of the tests.

Materials

The GORT-4 is an individually administered standardized norm-referenced test of oral reading for individuals ages 6 years 0 months through 18 years 11 months. The Examiner's Manual (Wiederholt & Bryant, 2001b) specifies that the purposes of the GORT-4 are to assist in the identification of students with reading problems, determine the oral text reading strengths and weaknesses of individuals, document an individual's progress in learning to read text, and aid researchers as a measurement tool in their studies. The test was normed

Table 2 Raw score performance of the native English speakers (non-ESL) and the speakers of English as a second language (ESL)

	Non-ESL (n=130)		ESL (n=63)	
Subtest	Mean	SD	Mean	SD
PPVT***	131.26	20.80	101.68	34.59
PIAT Spelling	69.25	14.51	67.94	15.60
BNT***	34.72	8.52	22.63	9.60
WJ Letter ID***	49.41	4.96	52.08	4.0
WJ Reading Flu.*	39.53	9.76	36.25	11.17
WJ Pass. Comp.***	25.67	4.35	20.71	4.29
WJ Word Attack***	13.55	6.38	17.97	5.96
TOLD Word Order***	9.84	4.83	5.90	3.92
Sight Word Effic.	63.80	11.97	62.67	12.42
Phonemic Decod. ***	18.05	10.84	31.19	11.96
CTOPP Elision	6.22	2.77	7.08	4.98
CTOPP Blend. Words	5.46	3.39	6.41	4.66
CTOPP Letter Naming[a]	33.47	9.37	36.13	10.88
CTOPP Color Naming[a]	54.56	16.55	59.35	23.21
GORT Rate	25.65	6.14	24.76	6.87
GORT Accuracy*	19.82	8.05	17.00	7.98
GORT Fluency	45.48	12.82	41.76	13.39
GORT Comprehension A	14.77	13.81	13.92	12.32
GORT Comprehension B	21.02	10.89	21.33	11.05

PPVT Peabody Picture Vocabulary Test, *PIAT Spelling* Peabody Individual Achievement Test Spelling, *BNT* Boston Naming Test, *WJ Letter ID* Woodcock Johnson III Letter Identification, *WJ Reading Flu.* Woodcock Johnson III Reading Fluency, *WJ Pass. Comp.* Woodcock Johnson III Passage Comprehension, *WJ Word Attack* Woodcock Johnson III Word Attack, *TOLD Word Order* Test of Language Development-Intermediate Word Ordering, *Sight Word Effic.* Test of Word Reading Efficiency Sight Word Efficiency, *Phonemic Decod.* Test of Word Reading Efficiency Phonemic Decoding Efficiency, *CTOPP Elision* Comprehensive Test of Phonological Processing Elision, CTOPP Blend. *Words* Comprehensive Test of Phonological Processing Blending Words, *CTOPP Letter Naming* Comprehensive Test of Phonological Processing Rapid Letter Naming, *CTOPP Color Naming* Comprehensive Test of Phonological Processing Rapid Color Naming, *GORT* Gray Oral Reading Tests, *GORT Comprehension A* GORT Comprehension subtest scored from story 1, *GORT Comprehension B* GORT Comprehension subtest scored from story 3

*p<0.05; ***p<0.001 for differences between the non-ESL and ESL groups

[a] Time in seconds

on 1,677 individuals in 28 states between the fall of 1999 and the fall of 2000. When compared to the percentages reported in 1997 by the US Bureau of the Census, the GORT-4 manual claims that the sample is representative of the United States in terms of geographic region, gender, race, rural/urban residence, ethnicity, income, educational attainment of parents, and disability (Wiederholt & Bryant, 2001b).

The GORT-4 measures four components of reading: Rate (the time taken by an individual to read each passage), Accuracy (the correct pronunciation of each word in the story), Fluency (the rate and accuracy scores combined), and Comprehension (the ability to answer the questions about each passage's content). Although scores for Rate, Accuracy, Fluency, and Comprehension are established, only the Fluency and Comprehension scores are used to determine basals and ceilings.

The GORT-4 has two parallel forms (Form A and Form B), each containing 14 progressively difficult passages. The tester provides the individual with a test book containing the passages, follows along on an examiner's copy of the passages, times the reader as he/she reads the passages aloud, and marks any mistakes the reader makes. Upon completion of reading each passage, the testers removes the passage, provides the individual with five multiple-choice reading comprehension questions, reads each question and answer option aloud, and records the individual's response choice. Administration takes between 15 and 30 min in this population.

The GORT-4 manual states that the sequence of the stories is based on several factors, including topics appropriate to various age levels, syntax complexity, vocabulary level, and the readability level according to the Flesch–Kincaid readability formula. The manual indicates that the stories get progressively more difficult based on these factors. Reliability of the subtests across both forms ranges from 0.85 to 0.99 for content sampling, test–retest, and interscorer differences. The manual provides both qualitative and quantitative analyses to indicate that the test has content, construct, and predictive validity (Wiederholt & Bryant, 2001b).

The GORT-4 is composed of stories taken from the GORT-3 with a rearrangement of the stories due to a new standardization of sample results. In addition, a new first story was added to the beginning of each of the two forms to help lower the measurement floor of the measures. The placement of GORT-3 stories in GORT-4 follows a consistent pattern from story 7 through story 14, with each passage in both forms of the GORT-4 being moved one place up; that is, GORT-3 story 6 became GORT-4 story 7; GORT-3 story 7 became GORT-4 story 8, etc. However, prior to GORT-4 story 7, the pattern of the placement of GORT-3 stories in GORT-4 appears less consistent. For example, GORT-3 story 1 became GORT-4 story 3 and GORT-3 story 3 became GORT-4 story 6. The change in placement and order was based on "new normative data" (Wiederholt, & Bryant, 2001b, p. x.).

According to the manual, the examiner selects the entry-level story for an individual based on either prior knowledge about the individual's general reading ability or based on the individual's grade level. Entry points are specified based on reading grade levels. Individuals at the grade 1 or 2 level begin with story 1, individuals at the grade 3 or 4 level begin with story 3, individuals at the grade 5 through 8 level begin with story 5, and individuals at the grade 9 through 12 begin with story 9. The manual specifies that for those with poor reading ability, a lower entry point should be initiated. Therefore, because of the nature of this study's sample, all participants started with story 1 instead of starting with story 3 or story 5. To our surprise, in the early stages of GORT- 4 administration, many of the participants in this study established a ceiling at story 1 or story 2. We therefore decided to continue to administer the passages in order until the next ceiling was reached. Therefore, all students were administered story 1, story 2, and story 3 of the GORT. Administration was discontinued when they established their ceiling at story 3 or later.

As mentioned previously, this study was conducted within a context of a larger study. Therefore, in addition to Form A of GORT-4 (Wiederholt & Bryant, 2001a), all participants were individually administered a battery of tests measuring word identification, fluency, comprehension, phonological processes, vocabulary, spelling, naming speed, and syntactic knowledge. Testing generally took approximately 2 h (occasionally up to 3 h) in one setting with breaks. The following tests were administered:

Woodcock Johnson Psycho-Educational Battery III (WJ-III; Woodcock et al., 2001) This test has been standardized on participants ages 2.0 through 90.0+. The following subtests were administered: Letter-Word Identification (reading a list words of graduated difficulty), Passage Comprehension (silently reading passages and completing cloze items), Word Attack

(reading a list of nonsense words of graduated difficulty), and Reading Fluency (reading in 3 min as many sentences as possible and deciding whether the statement is true or false).

Boston Naming Test (BNT; Kaplan, Goodlass, & Weintraub, 2001) This test has been designed for participants ages 5 through 79. Participants are asked to label drawings presented one at a time on cards.

Peabody Individual Achievement Test-Revised (PIAT-R) Spelling Subtest (Frederick & Markwardt, 1997) This test has been standardized on participants ages 5 through 22. Participants are asked to recognize correct word spellings in a multiple-choice format.

Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999) The version used in this study test has been standardized on participants ages 7 through 24. The following subtests were administered: Elision (repeating a word without a specific sound), Blending Words (combining sounds to make a whole word), Rapid Color Naming (naming presented colors as quickly as possible), and Rapid Letter Naming (naming presented letters as quickly as possible).

Peabody Picture Vocabulary Test-Third Edition (PPVT-III; Dunn & Dunn, 1988) This test has been standardized on participants ages 2 years 6 months through over 90.0 years old. Participants are provided with a template of pictures and are asked to select the picture which best represents the word stated by the tester.

Test of Word Reading Efficiency (TOWRE; Torgesen, & Wagner, 1999) This test has been standardized on participants ages 6 through 24. Both the Sight Word Reading Efficiency (reading a list of words aloud as quickly and accurately as possible in 45 s) and Phonemic Decoding Efficiency subtests (reading a list of nonwords aloud as quickly and accurately as possible in 45 s) were administered.

Test of Language Development-Intermediate, Third Edition (TOLD-I: 3; Hammill & Newcomer, 1997) This test has been standardized on participants ages 8 through 12.11 years of age. The Word Ordering subtest of the TOLD-I: 3 was administered. Participants are orally presented with a series of words and are asked to reorder these words and state them as a sentence.

The order of test administration was based on our previous experiences testing adult literacy students. We considered fatigue factors, test requirements, and practice effects when selecting the order. As a result, the tests were administered in the following order: WJ-III Letter-Word Identification, PPVT-III, PIAT Spelling, BNT, WJ-III Reading Fluency, WJ-III Passage Comprehension, WJ-III Word Attack, TOLD-I: 3 Word Ordering, GORT-4, TOWRE Sight Word Efficiency, TOWRE Phonemic Decoding Efficiency, CTOPP Elision, CTOPP Blending, CTOPP Rapid Letter Naming, and CTOPP Rapid Color Naming. For test performance on these assessments, see Table 2.

Results

Question 1: When starting with story 1, is the GORT a psychometrically appropriate test for adults who recognize words at the third to fifth grade levels?

The following analyses focused on results obtained when comprehension scores were obtained starting at story 1 (in the tables, denoted as Comprehension A).

Table 3 Intercorrelations between GORT subscales for the overall group, native English speakers (non-ESL), and speakers of English as a second language (ESL)

Subscale	1	2	3	4	5
Overall (n=193)					
Rate	–	0.63*		0.07	0.22*
Accuracy		–		0.12	0.26*
Fluency			–	0.11	0.26**
Comprehension A				–	−0.03
Comprehension B					–
Non-ESL (n=130)					
Rate	–	0.57*		−0.07	0.14
Accuracy		–		−0.10	0.19
Fluency			–	−0.10	0.19
Comprehension A				–	−0.03
Comprehension B					–
ESL (n=63)					
Rate	–	0.73**		0.18	0.39*
Accuracy		–		0.36	0.44*
Fluency			–	0.29	0.45*
Comprehension A				–	−0.04
Comprehension B					–

Comprehension A and B refer to the Comprehension subtest scored from story 1 and 3, respectively. Similar to the GORT-4 manual, the correlations between Fluency and both Rate and Accuracy are not provided because Fluency is a composite of Rate and Accuracy

*p<0.05; **p<0.01

Since the manual reports significant intercorrelations of the GORT-4 subtests to each other, our initial analyses focused on replicating these results with the adult literacy students. The manual reports correlations of 0.85 between Rate and Accuracy, while the ones found in this study are lower (see Table 3). In addition, the manual reports correlations of 0.39, 0.42, and 0.45 for Comprehension and Rate, Comprehension and Accuracy, and Comprehension and Fluency, respectively. Interestingly, the adult literacy sample's Comprehension scores when starting with story 1 (marked in the tables as Comprehension A) demonstrated no significant correlation with Rate, Accuracy, or Fluency (see Table 3).

Correlational analyses were next conducted to test the concurrent validity of the GORT-4 subtests with the battery of other measures used with the adult literacy students in this study. For both ESL and native speakers, intercorrelations of GORT-4 Rate, GORT-4 Accuracy, and GORT-4 Fluency with the other tests in the battery of assessments indicated significant relationships. Specifically, the GORT-4 Rate, Accuracy, and Fluency correlations with most of the other assessments ranged from 0.20 to 0.75 for native speakers of English and ranged from 0.25 to 0.78 for ESL participants. Some exceptions were noted. Specifically, the PPVT-III and BNT scores did not significantly correlate with GORT-4 Accuracy or Fluency subtest scores for either group. In addition, for both groups, the TOLD-I-3 Word Ordering subtest scores did not correlate significantly with the GORT-4 Rate subtest scores. Three differences in correlation patterns were noted across both groups: For native speakers, the TOLD 1–3 Word Ordering subtest scores were not significantly

correlated with the GORT-4 Accuracy and Fluency subtest scores, while they were significantly correlated for the ESL group (r=0.30 and r=0.28, respectively). Finally, for the native speakers, the GORT-4 Rate subtest scores were not correlated significantly for the PPVT-III, BNT, and CTOPP Blending Words subtest scores, while they were significantly correlated for the ESL group (r=0.25; r=0.28; r=0.28, respectively).

A very different pattern was noted for the correlation analyses for the GORT-4 Comprehension subtest scores. For both ESL and native English-speaking participants, compared to the Rate, Accuracy, and Fluency GORT-4 subtests, the GORT-4 Comprehension subtest was significantly correlated with fewer other tests in the battery. In addition, the magnitude of the significant correlations was small. Specifically, for the native group, the GORT-4 Comprehension A subtest was only significantly related to the PPVT-III (r=0.27), BNT (r=0.18), and TOLD-I: 3 Word Ordering (r=0.30). This is in contrast to the range of correlations (0.20 to 0.75) between the GORT-4 Rate, Accuracy, and Fluency scores with most of the other measures. The ESL group only showed a significant correlation of GORT-4 Comprehension A with the CTOPP Blending Words subtest (r=0.26).

Question 2: What happens when results are analyzed for those participants who reach a comprehension ceiling at story 1, and/or story 2, but establish a comprehension basal at or after story 3?

Although divergent patterns were not noted on GORT-4 Rate, Accuracy, or Fluency scores of the 193 participants, 50% (n=97) had an atypical pattern on their GORT comprehension performance; that is, they met the ceiling rule on story 1 and/or story 2, but established a basal at or after story 3 (hereafter this group is referred to as the atypical group). Specifically, as Table 4 indicates, 64% of the atypical participants (both native and non-native English-speaking participants) reached a ceiling at story 1. As part of the study, they continued to be administered additional passages, and after reading story 2, 62% of the atypical participants reached a ceiling. However, as part of the study procedure, they continued to be administered additional passages, and only 6% of these same participants attained a ceiling at story 3! Therefore, for the sample (both native and non-native English-speaking participants), the first and second stories were more difficult than the third story. This atypical pattern can be seen quite dramatically in Fig. 1, where more participants answered comprehension questions correctly in stories 3, 4, and 5 than in stories 1 and 2. As shown in Tables 1 and 5, with the exception of the TOLD-I:3 Word Ordering subtest, the participants who showed this atypical pattern were demographically similar and performed similarly on the reading battery of tests to individuals who did not show this atypical pattern.

Of the 97 participants who showed the atypical comprehension score pattern, 20 were between the ages of 16 years and 18 years 10 months. Similar to the entire sample, this group did not differ from the older participants in this group in their GORT-4 Rate, Accuracy, Fluency, and Comprehension scores. In addition, of the 97 participants with atypical GORT-4 Comprehension score patterns, 30 were ESL speakers and 67 were native speakers of English. No significant differences were found between the atypical ESL group and the atypical native speakers in terms of their GORT-4 Rate, Accuracy, Fluency, and Comprehension performances.

Table 6 indicates the intercorrelations of the GORT-4 subtests to each other for the atypical group. Similar to the findings reported under Question 1, the atypical group's Comprehension scores were not correlated with any of the other GORT-4 subtests when

Table 4 Number of atypical native English speakers (non-ESL) and atypical speakers of English as a second language (ESL) who read each story and the number who established a comprehension ceiling at each story

	Non-ESL (n=67)			ESL (n=30)		
Story	Reached ceiling	Read story[a]	% Reached ceiling	Reached ceiling	Read story	% Reached ceiling
1	43	67	64	19	30	63
2	42	67	63	18	30	60
1 and 2[b]	18	67	27	7	30	23
3	2	67	3	4	30	13
4	19	65	28	4	26	13
5	15	46	22	10	22	33
6	11	31	16	0	12	0
7	9	20	13	7	12	23
8	3	11	4	1	5	3
9	2	8	3	1	4	3
10	4	6	6	2	3	7
11	2	2	3	1	1	3
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0

[a] The number of students who read each story varied because at story 3 and beyond if a student established the ceiling, administration discontinued

[b] Some students established a ceiling on both story 1 and story 2

their scores started with story 1 (Comprehension A). However, when the scores were computed from a starting point of story 3 (Comprehension B), their Comprehension scores were significantly related to their GORT-4 Rate, Accuracy, and Fluency scores. This finding is more like the pattern found in the normative data reported in the GORT-4 manual; however, the correlations were not as large as those reported in the GORT-4 manual.

To examine the convergent validity of the GORT-4 Comprehension scores with the other reading and reading-related measures in the battery for the atypical pattern group, Pearson correlation coefficients were computed (see Table 7). When the ceiling rule was applied from story 1 (noted in the table as Comp A), there were no significant correlations between the GORT-4 Comprehension scores and any of the other tests. However, when the ceiling rule was applied from story 3 (i.e., giving full credit for stories 1 and 2 as indicated in the manual), there were significant correlations between the other tests and the GORT-4 Comprehension scores.

In an attempt to understand the underlying reason for the discrepancy, test performance was analyzed to ensure that the participants were actually able to read the passages well enough to comprehend them. Therefore, the first three stories were evaluated with regard to reading Rate and Accuracy. In terms of Rate, only two participants out of 97 (both native speakers) read the first story at what the GORT-4 considers a slow pace (i.e., taking 18 s or longer). None of the 97 atypical group participants read story 2 at a slow pace (i.e., taking 52 s or longer). Two participants out of the 97 (both ESL participants) read the third story at what the GORT-4 considers a slow pace (i.e., taking longer than 58 s), whereas all the native-speaking participants read the third story fast enough to move on to the next passage. Therefore, overall, rate was not an issue for the participants for the first three stories.

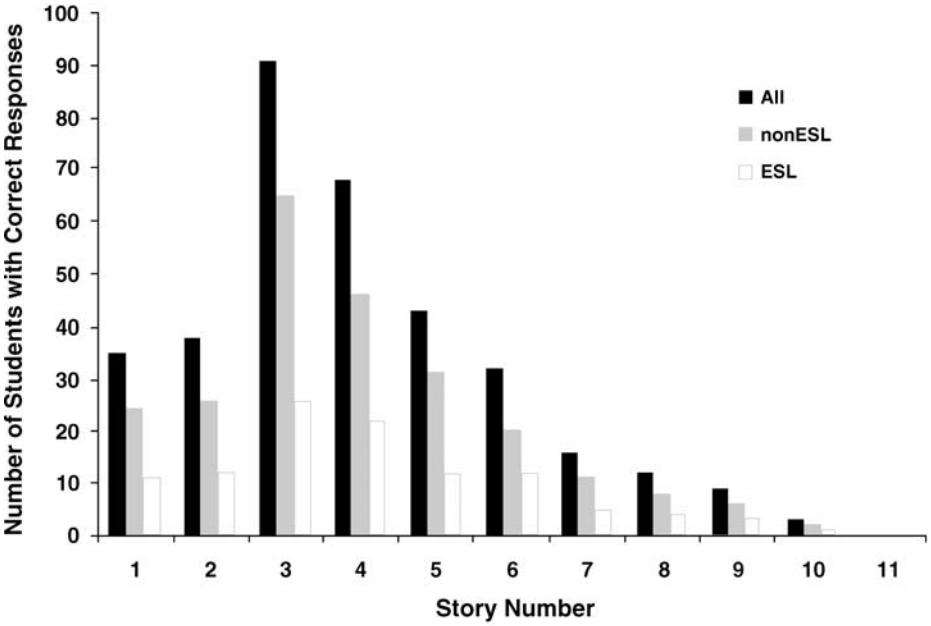

Fig. 1 The number of total students, native English-speaking students (***nonESL***), and students speaking English as a second language (*ESL*) with correct comprehension responses to each GORT story

With respect to Accuracy on story 1, nine native speakers and three ESL participants showed low accuracy (i.e., they made three or more deviations from print). On story 2, 19 native speakers and ten ESL participants showed low accuracy (i.e., they made six or more deviations from print). On story 3, 13 native speakers and seven ESL participants had low accuracy (i.e., they made five or more deviations from print). Therefore, compared to rate, accuracy of reading of the GORT-4 passages followed a more natural progression of difficulty, one that would be expected based on the readability index provided in the manual.

To analyze whether the atypical pattern existed when individuals were grouped by their scores on another comprehension test, the participants' WJ-III Passage Comprehension (Woodcock et al., 2001) scores were used to group participants. As indicated in Table 8, both native English speakers and ESL participants experienced more difficulty with the first and second GORT-4 stories than they did on the third GORT-4 story regardless of WJ-III Passage Comprehension grade level. Specifically, participants who scored below the third grade level equivalency on the WJ-III Passage Comprehension subtest as well as those who scored at or above the third grade level equivalency had more difficulties with stories 1 and 2 than they had with story 3.

Given the fact that so many of the participants had difficulty with the first two GORT-4 passages, it was questioned whether these atypical pattern results were due to their initial test anxiety. In other words, perhaps due to test anxiety, the participants performed poorly on the initial items of the test, only to become comfortable and perform better on items administered later in the test. To evaluate this possibility, participant performance on the first 15 items of the WJ-III Passage Comprehension subtest (Woodcock et al., 2001) was examined. On this subtest, if a participant correctly answers the first 15 items, he/she is considered to have a third grade comprehension level. Results indicated that fewer than 10% of the sample failed any of the first 15 items. This suggests that the atypical

Table 5 Subtest raw scores for the overall group and a comparison of subtest raw scores for the atypical pattern group and the non-atypical pattern group

	Overall (n=193)		Atypical (n=97)		Non-atypical (n=96)	
Subtest	Mean	SD	Mean	SD	Mean	SD
PPVT	121.61	29.50	119.29	30.47	123.95	28.47
PIAT Spelling	68.82	14.85	69.14	14.68	68.50	15.09
BNT	30.76	10.53	30.29	10.43	31.23	10.67
WJ Letter ID	50.28	4.82	50.26	4.43	50.30	5.21
WJ Reading Fluency	38.46	10.33	37.41	11.00	39.52	9.54
WJ Pass. Comprehension	24.05	4.91	23.57	5.07	24.54	4.70
WJ Word Attack	14.99	6.57	15.10	6.66	14.89	6.51
TOLD Word Order*	8.55	4.90	7.80	4.63	9.31	5.08
Sight Word Efficiency	63.43	12.10	63.37	12.75	63.49	11.48
Phonemic Decoding	22.29	12.77	22.67	12.82	21.91	12.78
CTOPP Elision	6.49	3.64	6.02	3.64	6.98	3.60
CTOPP Blending Words	5.76	3.85	5.25	3.72	6.28	3.94
CTOPP Letter Naming[a]	34.34	9.94	34.77	9.90	33.90	10.02
CTOPP Color Naming[a]	56.11	19.03	56.57	15.72	55.64	21.98
GORT Rate	25.36	6.38	25.26	6.87	25.47	5.88
GORT Accuracy	18.90	8.12	17.97	7.97	19.84	8.19
GORT Fluency	44.27	13.09	43.23	13.39	45.32	12.76
GORT Comprehension A***	14.49	13.31	3.05	2.25	26.17	9.06
GORT Comprehension B	21.11	10.89	21.11	10.89	–	–

PPVT Peabody Picture Vocabulary Test, *PIAT Spelling* Peabody Individual Achievement Test Spelling, *BNT* Boston Naming Test, *WJ Letter ID* Woodcock Johnson III Letter Identification, *WJ Reading Flu.* Woodcock Johnson III Reading Fluency, *WJ Pass. Comp.* Woodcock Johnson III Passage Comprehension, *WJ Word Attack* Woodcock Johnson III Word Attack, *TOLD Word Order* Test of Language Development-Intermediate Word Ordering, *Sight Word Effic.* Test of Word Reading Efficiency Sight Word Efficiency, *Phonemic Decod.* Test of Word Reading Efficiency Phonemic Decoding Efficiency, *CTOPP Elision* Comprehensive Test of Phonological Processing Elision, CTOPP Blend, *Words* Comprehensive Test of Phonological Processing Blending Words, *CTOPP Letter Naming* Comprehensive Test of Phonological Processing Rapid Letter Naming, *CTOPP Color Naming* Comprehensive Test of Phonological Processing Rapid Color Naming, *GORT* Gray Oral Reading Tests, *GORT Comprehension A* GORT Comprehension subtest scored from story 1, *GORT Comprehension B* GORT Comprehension subtest scored from story 3

*p<0.05; ***p<0.001 for differences between the atypical and non-atypical groups

[a] Time in seconds

performance on the GORT-4 Comprehension was probably not due to test anxiety. However, this possibility begs empirical testing, as the WJ-III Passage Comprehension subtest (Woodcock et al., 2001) does not require participants to read aloud, while the GORT does.

An alternate explanation is that the participants became bored with the first two passages in the GORT and thus did not pay adequate attention to passage content. Rather than anxiety provoking, perhaps the stories were simply too easy and insufficiently engaging for the participants. Unfortunately, participants were not questioned about their feelings about the GORT passages, and therefore, this possibility awaits future study.

Table 6 Intercorrelations of GORT subtests for the group with atypical patterns (n=97)

Subtest	1	2	3	4	5
Rate	–	0.63*		0.00	0.22*
Accuracy		–		0.01	0.26*
Fluency			–	0.01	0.26**
Comprehension A				–	−0.03
Comprehension B					–

Comprehension A comprehension subtest scored from story 1, *Comprehension B* comprehension subtest scored from story 3

*p<0.05; **p<0.01

Discussion

Due to the paucity of tests developed for adult literacy students, most assessments used in research with adult literacy students are tests standardized on children and therefore associated with typical childhood developmental units (Sabatini, et al., 2000). This study addresses the complexities and problems of using a test normed on children to measure reading comprehension skills of adults who have severe difficulty reading. Effectively measuring adults' reading performance in a valid and discriminating way is important, and this study suggests that educators and researchers should be very careful when interpreting test results of adults who have difficulty reading when children's norm-referenced tests are administered.

The complexities addressed in this paper are similar to "out-of-level testing" issues that arise when children in special education are assessed. Out-of-level testing refers to "... the practice of testing a student who is in one grade with a level of a test developed for students in either one or more grades above or below that grade." (Minnema, Thurlow, Bielinski, & Scott, 2000 p. 1). There is more than a 30-year precedence of testing students out of level. Examples of reasons that are often given are that students who cannot read the test items will get credit for guessing correct answers and that students may find it emotionally difficult to take a test that is harder than their ability level. Many argue that if low performing students start at a lower level, they are probably better-matched in terms of their ability level to the difficulty level of the test (Minnema et al., 2000). However, as Bielinski, Thurlow, Minnema, and Scott (2000) indicate, the raw scores of students who have been tested out of level do not possess the same precision of measurement as the raw scores of students who do start at the manual's starting level. As Salvia and Ysseldyke (2004) state, "... a person's performance on a test is measured in reference to the performances of others who are presumably like that person in other respects. " (p. 30). When this done not happen, there is no common scale. This is similar to what was found in this study.

In order to increase the comfort level during reading tests of adults who have difficulty reading, administration of tests are often started at a lower level than is typically recommended. This philosophy is in agreement with the GORT-4 manual (Wiederholt & Bryant, 2001b), which clearly specifies that a lower entry point should be employed when testing those with poor reading abilities. Therefore, in this study, all participants started with

Table 7 Correlations of GORT comprehension scores (Ceiling Rules Applied at story 1 and story 3) with other subtests for students with atypical patterns

	Total (n=97)		Non-ESL (n=67)		ESL (n=30)	
Subtest	Comp A	Comp B	Comp A	Comp B	Comp A	Comp B
PPVT	0.08	0.40**	0.09	0.48**	0.08	0.40*
PIAT Spelling	−0.02	0.18	−0.05	0.10	0.04	0.32
BNT	0.02	0.44**	0.03	0.55**	−0.03	0.46*
WJ Letter ID	0.00	0.24*	−0.02	0.28*	0.12	0.15
WJ Reading Flu.	−0.07	0.28**	−0.10	0.21	0.01	0.42*
WJ Pass. Comp.	−0.06	0.42**	−0.11	0.42**	0.05	0.57**
WJ Word Attack	−0.04	0.15	−0.06	0.15	0.00	0.14
TOLD Word Order	0.15	0.41**	0.18	0.48**	0.09	0.33
Sight Word Effic.	−0.13	0.23*	−0.11	0.15	−0.20	0.40*
Phonemic Decod.	−0.15	0.06	−0.22	0.04	−0.01	0.10
CTOPP Elision	−0.02	0.29*	−0.01	0.27*	−0.05	0.34
CTOPP Blend. Words	0.13	0.16	0.06	0.01	0.30	0.41*
CTOPP Letter Naming	0.17	−0.09	0.19	0.01	0.12	−0.38*
CTOPP Color Naming	0.03	0.08	0.04	0.12	0.01	0.02
GORT Rate	0.00	0.22*	−0.07	0.14	0.18	0.39*
GORT Accuracy	0.01	0.26*	−0.10	0.19	0.36	0.44*
GORT Fluency	0.01	0.26**	−0.10	0.19	0.29	0.45*
GORT Comp A	–	−0.03	–	−0.03	–	−0.04
GORT Comp B	−0.03	–	−0.03	–	−0.04	–

PPVT Peabody Picture Vocabulary Test, *PIAT Spelling* Peabody Individual Achievement Test Spelling, *BNT* Boston Naming Test, *WJ Letter ID* Woodcock Johnson III Letter Identification, *WJ Reading Flu.* Woodcock Johnson III Reading Fluency, *WJ Pass. Comp.* Woodcock Johnson III Passage Comprehension, *WJ Word Attack* Woodcock Johnson III Word Attack, *TOLD Word Order* Test of Language Development-Intermediate Word Ordering, *Sight Word Effic.* Test of Word Reading Efficiency Sight Word Efficiency, *Phonemic Decod.* Test of Word Reading Efficiency Phonemic Decoding Efficiency, *CTOPP Elision* Comprehensive Test of Phonological Processing Elision, CTOPP Blend, *Words* Comprehensive Test of Phonological Processing Blending Words, *CTOPP Letter Naming* Comprehensive Test of Phonological Processing Rapid Letter Naming, *CTOPP Color Naming* Comprehensive Test of Phonological Processing Rapid Color Naming, *GORT* Gray Oral Reading Tests, *GORT Comp A* GORT Comprehension subtest scored from story 1, *GORT Comp B* GORT Comprehension subtest scored from story 3

*p<0.05; **p<0.01

story 1, which is the recommended entry level for students in Grades 1 and 2. By doing so, two results are worthy of reporting:

- Comprehension subtest scores were not correlated with Accuracy, Rate, and Fluency subtests scores.
- The Comprehension subtest was very poorly correlated with other reading-related tests.

When scores were analyzed based on starting the test on story 3, the typical starting point for readers at the third and fourth grade levels, the following was noted:

- Comprehension subtest scores were correlated with Accuracy, Rate, and Fluency subtest scores.

Table 8 Number of atypical native speakers of English (non-ESL) and atypical speakers of English as a second language (ESL) reading GORT stories and reaching ceiling at each story by Woodcock Johnson Passage Comprehension Grade Equivalency Groupings

	Non-ESL (n=67)				ESL (n=30)			
	WJ PC GE <3.0 (n=38)		WJ PC GE ≥3.0 (n=29)		WJ PC GE <3.0 (n=24)		WJ PC GE ≥3.0 (n=6)	
Story	Reached ceiling	Read story	Reached ceiling	Read story	Reached ceiling	Read story	Reached ceiling	Read story
1	25	38	18	29	16	24	3	6
2	25	38	17	29	14	24	4	6
3	2	38	0	29	3	24	1	6
4	12	36	7	29	4	21	0	5
5	13	24	2	22	8	17	2	5
6	8	11	3	20	0	9	0	3
7	0	3	9	17	6	9	1	3
8	0	3	3	8	1	3	0	2
9	1	3	1	5	1	2	0	2
10	1	2	3	4	1	1	1	2
11	1	1	1	1	0	0	1	1
12	0	0	0	0	0	0	0	0

The number of students who read each story varied because at story 3 and beyond if a student established the ceiling, administration discontinued

WJ PC GE Woodcock Johnson III Tests of Achievement, Passage Comprehension subtest, Grade Equivalency

- The Comprehension subtest was correlated with other tests related to reading.

More participants answered comprehension questions correctly about stories 3, 4, and 5, than about stories 1 and 2. In fact, 50% of the sample established a ceiling at story 1 but proceeded to establish a basal at or after story 3. This result was found regardless of participants' age, gender, educational attainment level, native English-speaking status, and WJ-III Passage Comprehension scores. If the manual had been followed and had test administration started on story 3 (for the participants who identified words at the third and fourth grade levels), or story 5 (for the participants who identified words at the fifth grade levels), many of the participants would have been awarded full credit below their highest basal, when in fact this study now provides data, indicating that this basal scoring assumption is not valid on the GORT-4 for the adult literacy population.

This study points to the complexity of scaling. As Crocker and Algina (1986) state, scaling involves that "... a hypothesis has been implicitly formulated that the construct is a property occurring in varying amounts so that it can be quantified using the proposed scaling rule on a theoretical unidimensional continuum" (p. 49). "The total score is computed by simply assuming that point values assigned to each possible response form a numeric scale... with order and equal units." (p. 50). As indicated by the results of this study, these underlying implications cannot be assumed with the GORT-4 and adult literacy students.

Does the GORT-4 have face validity for adult literacy students? The GORT-4 manual (Wiederholt & Bryant, 2001b) states: "The story format has tremendous face validity in that

it is... commonly used in classroom assessment and everyday life..." (p. 72). This is true for children, but not necessarily for adult literacy students. According to the manual, stories were selected for interest value. Although the stories may be interesting to children, they may not be for adults. As the GORT-4 manual (Wiederholt & Bryant, 2001b) mentions, other tests use the story format. In a future study, it would be interesting to see how the adult learners do on those other tests. Is there something peculiar about reading passages and being asked to answer questions about them?

A question unanswered by this study is whether the reported findings are only unique to adults who recognize words between the third and fifth grade levels. Therefore, future research should replicate this study with adult readers at all levels (first grade through expert levels), as well as with typically developing children. In addition, the ESL participants in this study were very diverse and yet were treated homogenously in the data analyses. Future researchers should specifically focus on ESL participants and analyze whether the findings in this study are replicated with a larger sample of diverse ESL participants. If similar findings are uncovered with typically developing children, adults with varying reading levels, and ESL students with varying backgrounds, then it is clear that the problems are not specific to adults with extreme reading difficulties, but rather the problems are specific to the test. Keenan and Betjemann's (2006) study with children provides some support for this possibility. They found that a high number of GORT Comprehension items could be answered by children based on their prior knowledge, without reading the passage.

Finally, an error analysis of each GORT item may yield interesting results. The manual lists four types of questions that are utilized in the test: literal, inferential, critical, and affective. Unfortunately, the manual does not label the questions, but possibly adult literacy students have more difficulty with certain question types over others. For example, Dewitz and Dewitz (2003) analyzed the comprehension errors of nine children and found that errors varied in terms of relational inferences, causal inferences, elaboration, syntax, and vocabulary. It would be interesting if adults who have difficulty reading showed variations in their comprehension error types.

In conclusion, this study points to the importance of research in the use of commercially standardized tests on special populations. Specifically, research on test characteristics of adult literacy students is needed not only by adult literacy researchers but also by all individuals charged with measuring the reading skills of adults. Clearly, more tests are needed that are specifically normed on low literate adults. Test users should be cautious in test result interpretation when administering the GORT to adults with reading difficulties.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River: Prentice Hall.

Bielinski, J., Thurlow, M., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores*. Minneapolis: National Center on Educational Outcomes. ERIC Document Reproduction Service No. ED449174.

Costenbader, V. K., & Adams, J. W. (1991). A review of the psychometric features of the PIAT-R: Implications for the practitioner. *Journal of School Psychology, 29*, 219–228. doi:10.1016/0022-4405(91)90003-A.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart, & Winston.

D'Amico-Samuels, D. (1991). *Perspectives on assessment from the New York City Adult Literacy Initiative: A critical issues paper*. New York: New York Literacy Assistance Center. ERIC Document Reproduction Service No. ED357658.

Dewitz, P., & Dewitz, P. K. (2003). They can read the words, but they can't understand: Refining comprehension assessment. *The Reading Teacher, 56*, 422–435.

Dunn, L. M., & Dunn, L. M. (1988). *Peabody Picture Vocabulary Test—Revised*. Circle Pines: American Guidance Service.

Frederick, M., & Markwardt, T. (1997). *Peabody Individual Achievement Test—Revised*. Circle Pines: American Guidance Service.

Greenberg, D., Ehri, L., & Perin, D. (1997). Are word reading processes the same or different in adult literacy students and 3rd–5th graders matched for reading level? *Journal of Educational Psychology, 89*, 262–275. doi:10.1037/0022-0663.89.2.262.

Greenberg, D., Ehri, L. C., & Perin, D. (2002). Do adult literacy students make the same word-reading and spelling errors as children matched for word-reading age? *Scientific Studies of Reading, 6*, 221–243. doi:10.1207/S1532799XSSR0603_2.

Hammill, D. D., & Newcomer, P. L. (1997). *Test of language development—Intermediate* (3rd ed.). Austin: Pro-Ed.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test*. Baltimore: Lippincott Williams & Wilkins.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10*, 363–380. doi:10.1207/s1532799xssr1004_2.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., & Dunleavy, E. (2007). Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy. National Center for Education Statistics. National Assessment of Adult Literacy (NAAL): A first look at the literacy of America's adults in the 21st century. Available at: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007480 Accessed April 10, 2007

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis*. Minneapolis: National Center on Educational Outcomes. ERIC Document Reproduction Service No. ED446409.

Patterson, M. B. (2008). Learning disability prevalence and adult education program characteristics. *Learning Disabilities Research & Practice, 23*, 50–59.

Perin, D. (1991). Test scores and adult literacy instruction: Relationship of reading test scores to three types of literacy instruction in a worker education program. *Language and Literacy Spectrum, 1*, 46–51.

Sabatini, J., Venezky, R.L., Kharik, P., & Jain, R. (2000). Cognitive reading assessment for low literate adults: An analytic review and new framework. Technical Report, October 9, 1995. National Center on Adult Literacy, Philadelphia, PA. (ERIC Document Reproduction No. ED447308).

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9th ed.). New York: Houghton Mifflin.

Stevens, K. B., & Price, J. R. (1999). Adult reading assessment: Are we doing the best with what we have? *Applied Neuropsychology, 6*, 68–78. doi:10.1207/s15324826an0602_2.

Torgesen, J., & Wagner, R. K. (1999). *Test of Word Reading Efficiency*. Austin: Pro-Ed.

Venezky, R. L., & Sabatini, J. P. (2002). Introduction to this special issue: Reading development in adults. *Scientific Studies of Reading, 6*, 217–220. doi:10.1207/S1532799XSSR0603_1.

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive test of phonological processes*. Austin: Pro-Ed.

Wiederholt, J. L., & Bryant, B. R. (2001a). *Gray Oral Reading Test-IV.* Austin: Pro-Ed.

Wiederholt, J. L., & Bryant, B. R. (2001b). *Gray Oral Reading Test-IV Examiner's Manual*. Austin: Pro-Ed.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca: Riverside.