DOI 10.3758/s13428-014-0534-3

Reading ability and print exposure: item response theory analysis of the author recognition test

Mariah Moore · **Peter C. Gordon**

Published online: 20 November 2014

Abstract In the author recognition test (ART), participants are presented with a series of names and foils and are asked to indicate which ones they recognize as authors. The test is a strong predictor of reading skill, and this predictive ability is generally explained as occurring because author knowledge is likely acquired through reading or other forms of print exposure. In this large-scale study (1,012 college student participants), we used item response theory (IRT) to analyze item (author) characteristics in order to facilitate identification of the determinants of item difficulty, provide a basis for further test development, and optimize scoring of the ART. Factor analysis suggested a potential two-factor structure of the ART, differentiating between literary and popular authors. Effective and ineffective author names were identified so as to facilitate future revisions of the ART. Analyses showed that the ART is a highly significant predictor of the time spent encoding words, as measured using eyetracking during reading. The relationship between the ART and time spent reading provided a basis for implementing a higher penalty for selecting foils, rather than the standard method of ART scoring (names selected minus foils selected). The findings provide novel support for the view that the ART is a valid indicator of reading volume. Furthermore, they show that frequency data can be used to select items of appropriate difficulty, and that frequency data from corpora based on particular time periods and types of texts may allow adaptations of the test for different populations.

Keywords Print exposure · Author recognition test · Item response theory · Eye movements · Reading

M. Moore · P. C. Gordon (✉)
Department of Psychology, CB#3270, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3270, USA
e-mail: pcg@email.unc.edu

Imagine a test that simply asks whether J. R. R. Tolkien and Kurt Vonnegut are names of authors. It is not obvious that testing these seemingly arbitrary bits of culturally specific knowledge should tell us anything about the operation of basic cognitive processes. However, beginning with the work of Stanovich and West (1989), author and literature recognition tests have been found to provide a very efficient and surprisingly effective way of predicting a number of component reading skills across a range of age groups (Mol & Bus, 2011). Stanovich and West argued that knowledge of authors and literature is predictive of reading skill because such knowledge is most likely to be acquired through reading, and therefore can serve as an indication of an individual's reading volume or print exposure. Of course, the evidence that knowledge of authors and literature reflects print exposure is necessarily indirect. Furthermore, because this knowledge is culturally specific, the items on the test must be chosen so that they are appropriate for the population being studied.

The author recognition test (ART) is a list of author and nonauthor names, with participants simply being asked to check those names that they recognize to be authors. The ART consistently outperforms other literature recognition tests, such as magazine or newspaper recognition tests, as a predictor of reading-related variables such as spelling ability, word recognition, and cultural literacy (Stanovich & West, 1989; West, Stanovich, & Mitchell, 1993), and it has been found to have high reliability (α = .84 in Stanovich & West, 1989; Mol & Bus, 2011, reviewed reports of α = .75–.89). Questionnaires and diary studies provide alternative methods for assessing reading volume that seem to be more direct than the ART. However, self-report reading questionnaires introduce bias in the form of inflated scores, since reading is considered socially desirable (Stanovich & West, 1989). The ART circumvents the tendency to provide socially desirable answers when self-reporting how often one reads. Diary studies can provide valid measures of print exposure, but the sustained effort that they require from participants leads to attrition, whereas in contrast the ART is nonintrusive and quicker (Carp & Carp, 1981; Mol & Bus, 2011). Versions of the ART and of related book title recognition tests (i.e., the Children's Title Checklist or Children's Author Checklist) have been found to successfully predict vocabulary and word recognition in children (Cunningham & Stanovich, 1990; Sénéchal, LeFevre, Hudson, & Lawson, 1996).

The original ART aimed to measure extracurricular reading, and for that reason it consisted mostly of the names of authors of popular fiction (Stanovich & Cunningham, 1992). Validation of its success in measuring extracurricular reading has come from a variety of types of evidence: West et al. (1993) approached individuals who were waiting in airport terminals and asked them to complete the ART; those who had been reading when they were approached scored higher than those who had not been reading. Students who indicate a preference for reading as a leisure activity, as compared to music or television, also score higher on the ART (Mol & Bus, 2011). Finally, ART scores correlate with self-reported time spent reading, particularly for fiction (Acheson, Wells, & MacDonald, 2008).

Although it has been shown that personally having read books by the authors in the ART correlates more strongly to reading ability than does simply knowing the authors without having read their books (Martin-Chang & Gould, 2008), the ART is not based on the assumption that recognizing an author's name implies having read the author's work. Instead, it assumes that reading a lot increases the chance of having a level of exposure to the authors that is sufficient to recognize their names. Thus, although the ART directly tests a particular type of knowledge, its effectiveness is not thought to solely depend on that knowledge per se, but instead depends on how that knowledge reflects differential practice at reading. Differences in practice affect reading skill, which in turn affects the degree to which reading is a rewarding or unrewarding experience. This may lead to a "causal spiral," in which print exposure stimulates reading development, which in turn makes reading more rewarding and leads to more reading. Thus, the rich get richer, while the poor fall further behind ("the Matthew effect"; Stanovich, 1986), with the amount of variance in reading skill that is accounted for by print exposure increasing from elementary school to middle school, high school, and college (Mol & Bus, 2011).

At a general level, ART scores relate to academic achievement and IQ with a small effect size (Mol & Bus, 2011). In terms of more specific abilities, ART scores correlate moderately with vocabulary knowledge (.60 in Stanovich & Cunningham, 1992; .54 in Beech, 2002), with the ART accounting for additional variance in vocabulary, verbal comprehension, and general ability after nonverbal ability has been taken into account (Stanovich & Cunningham, 1992; West et al., 1993). ART scores also account for unique individual variation in word identification, word naming, and spelling after phonological and orthographic processing have been accounted for (Stanovich & West, 1989). In addition to their relation to self-reports of the amount of time spent reading, ART scores are also related to self-reports of reading speed (Acheson et al., 2008). This relation to reading speed is corroborated by experimental results showing that low ART scores are associated with longer times and greater word frequency effects in lexical decision tasks (Chateau & Jared, 2000; Sears, Campbell, & Lupker, 2006) and in studies using eyetracking during normal reading. In eyetracking studies, the most common measure of the time required for word recognition is called *gaze duration*; it is the sum of the durations of first-pass fixations on a word, and it is very sensitive to word frequency and predictability (Inhoff & Rayner, 1986; Rayner, 1998; Rayner & Duffy, 1986). Higher ART scores are associated with shorter gaze durations, as well as with reductions in the effect of word frequency on gaze duration (i.e., high word frequency does not raise gaze duration times as much for high ART scorers; Gordon, Moore, Choi, Hoedemaker, & Lowder, 2014; Sears et al., 2006).

The present study had two major goals. At a theoretical level, it provided a novel test of the idea that the likelihood of recognizing an author on the ART is based on having encountered that author's name while reading. It does so by examining whether variation in the difficulty of author items on the ART is related to the frequency with which the authors' names appear in print. At a practical level, we used item response theory (IRT) to gain a better understanding of the psychometric strengths and weaknesses of the ART. Previous versions of the ART have been scored by taking the number of authors correctly selected minus the number of foils (nonauthors) incorrectly selected (Acheson et al., 2008; Martin-Chang & Gould, 2008; Stanovich & West, 1989). As such, they have relied implicitly on classical test theory, in which the sum of item responses is taken as an estimate of true scores. In contrast, IRT is a model-based approach that takes into account how each item performs at different levels of participant ability in a way that allows for more accurate measures of internal consistency and for an assessment of more complex patterns of data (Steinberg & Thissen, 1996). IRT is particularly valuable in identifying which items should be retained or eliminated from the test and in facilitating the creation of new versions of tests with scores that can be understood in relation to scores on earlier versions. Because the ART is linked to a particular culture and point in time, it is important to have effective procedures for creating new versions. These theoretical and practical goals were pursued using ART data from a large sample of college students. In addition, data about eye movements during reading were available for most of these participants, and results on a measure of processing speed (rapid automatized naming; Denckla & Rudel, 1974) and on a vocabulary test were available for a smaller number of participants. The efficiency of word recognition, as indicated by gaze durations from the eyetracking data, was used as an external criterion variable for validating the psychometric evaluation of the ART.

Method

Participants

A total of 1,012 students at the University of North Carolina at Chapel Hill participated in exchange for credit in an Introduction to Psychology course. All of the participants were native English speakers with normal or corrected-to-normal vision. Approximately 60.6 % of the participants were female; 76 % were White, 11.3 % were Black or African American, and 6.3 % were Asian. About 5.5 % of the participants were Hispanic or Latino.

Procedure

Individual-differences data were collected as part of a series of 23 experiments, each of which was designed to test specific psycholinguistic hypotheses. In addition to the ART, individual-differences measures were obtained in a short vocabulary test (205 participants) and the rapid automatized naming (RAN) test (569 participants). Twenty-one of the psycholinguistic experiments (testing 789 participants) provided recordings of eye movements during sentence reading; data from those studies were used to assess the relationship between ART performance and reading skill in the present study. The remaining psycholinguistic experiments either did not involve recording eye movements or did not involve sentences; the data from those experiments were used here only to examine the psychometric properties of the ART. Each experimental session began with the psycholinguistic experiment, which was followed by administration of the individual-differences tasks. Data were collected over a period of approximately 4 years (fall 2010 through spring 2014).

Author recognition test Each participant completed an ART that used the 65 author names from the Acheson et al. (2008) version of the test along with 65 additional names that did not refer to known authors. The nonauthor foils were taken from the Martin-Chang and Gould (2008) adaption of the Stanovich and West (1989) ART. All names were listed in alphabetical order by last name. The test was administered by paper, and participants were asked to circle the names that they recognized as those of authors, but they were warned that their score would be penalized for circling nonauthors. An individual administration of the ART typically took around 3 min.

Vocabulary test A short (16-item) vocabulary test, based on the 14-item Wordsumplus, was administered (Cor, Haertel, Krosnick, & Malhotra, 2012) to 205 of the participants. Two of the easiest words on the Wordsumplus were dropped after a pilot study showed that almost all of the participants correctly responded to them; four additional difficult items that were not previously on the test were added.[1]

Rapid automatized naming (RAN) In the RAN task (Denckla & Rudel, 1974), participants are presented with an array of 36 items arranged on a four by nine grid and are asked to read or name them out loud as quickly as possible without making mistakes. There were two trials of each of the four RAN types (objects, colors, letters, and digits), and each participant's final score was computed as the average completion time across all eight trials.

Eyetracking experiments Twenty-one eyetracking experiments tested 23 to 52 participants each as they read from 28 to 152 sentences. The individual experiments consisted of sentences that manipulated a variety of factors, such as Lexical Repetition, Orthographic Neighborhood Size, Animacy of Words, Syntactic Structure, and Word Predictability. Eye movements were recorded from the participant's dominant eye using an SR EyeLink 1000. Stimuli were presented on a 20-in. ViewSonic G225f Monitor at a distance of 61 cm with a display resolution of 1,024 × 768. Before each session, the tracker was calibrated using a 9-point procedure; the calibration was checked between trials and the tracker was recalibrated when necessary. A chin-and-forehead rest minimized the head movements of participants. Each trial began with a fixation point on the left side of the screen on the horizontal axis. Once this point was fixated, the next screen displayed the sentence. Participants were instructed to read each sentence silently at a natural pace and then to press a button on a handheld console. They then answered a true-or-false question. The experimenter monitored eye movements throughout the session. Each eyetracking experiment began with four warm-up sentences.

Gaze duration was calculated as the sum of the durations of the first-pass fixations on a word until the eyes moved away from the word, either to the right or the left. The first two and final two words in a line were excluded from the analyses, as were function words and words with four or fewer letters. These exclusions were made so that gaze duration would not be influenced by the preparation and execution of long saccades at the ends of lines or by the high skipping rates

[1] Items B and F were dropped from the Wordsumplus. Items 17, 20, 21, and 27 were added from the Modified Vocabulary Test (Mayer, Panter, & Caruso, 2012). These changes to the Wordsumplus represented preliminary efforts to assess the population of interest with a short vocabulary test. It should not be viewed as a final test, but as an initial method of establishing the convergent validity of the ART.

observed for function words and short words. Word frequencies for the included words were calculated using SUBTLEX-US (Brysbaert & New, 2009).

Results

ART scores and items

The standard method of scoring the ART is to subtract the number of false alarms (foils incorrectly selected) from the number of hits (the number of authors correctly selected). Table 1 lists the means, standard deviations, and ranges of performance as assessed with the standard method (standard ART score) and the number of authors correctly selected without a penalty for selecting foils (name score). The other scoring methods will be discussed after the factor analysis is presented. Performance varied greatly across participants; test scores had a positive skew, and the maximum score of 46 out of 65 indicated the high difficulty of the test.

Table 2 shows the selection rates for individual author names. The correct response rates for individual author names ranged from 92.2 % for Ernest Hemingway to 0.3 % for Bernard Malamud. The mean selection rate was 23.8 %. The mean number of errors per participant was 0.74, with a standard deviation of 1.64. Four of the foils were never selected, and 8.9 % of the participants selected the most alluring foil (Mark Strauss). Two foil names, Mark Strauss and Robert Emery, were selected 2.5 times above the mean rate of foil responses, likely due to their similarity to author names or to the existence of authors with similar names. For example, Robert Emery may have been mistaken for bestselling author Robert Ellis. Table 3 shows participants' selections of foils, which were low overall.

Relation to other ART data, vocabulary, RAN, and sentence comprehension

Seventeen items in the Acheson et al. (2008) test came from the items included in two of the early versions of the ART (Stanovich & Cunningham, 1992; Stanovich & West, 1989). The mean selection rates aggregated from those Stanovich studies did not correlate significantly with our selection rates, $r(17) = .27$, n.s. However, the mean selection rates for those items in the Acheson et al. study were highly correlated with our rates, $r(17) = .87$, $p < .001$ [all Acheson study items correlated similarly, $r(65) = .88$, $p < .001$]. Our mean selection was lower than in the Acheson et al. study (24 %, as compared to 36 %). We will further compare the Acheson and Stanovich selection rates when discussing author frequencies.

Table 1 Author recognition test (ART) results with different scoring methods: Numbers of items, means, standard deviations, and ranges

	Scales	*N*	*M*	*SD*	Range
65-author scale	Standard ART score	1,012	14.72	7.32	48.0
	ART name score	1,012	15.47	7.50	50.0
50-author scale	Standard 50 ART score	1,012	13.75	6.81	44.0
	50 ART name score	1,012	14.49	6.88	46.0
	50 IRTScore – 2 errors	1,012	13.01	7.14	55.5

We found a moderate correlation between the standard ART score and our modified Wordsumplus vocabulary test, $r(205) = .44$, $p < .001$. There was a statistically significant but very small correlation between standard ART scores and average completion times on the RAN, $r(569) = –.09$, $p = .044$. Although RAN performance predicts an impressive range of reading abilities (Kirby et al., 2010), for college students it appears to be largely unrelated to the ART. More information on how RAN performance relates to reading in college students can be found in Gordon et al. (2014) and in Gordon and Hoedemaker (2014). There was a very small correlation between standard ART scores and accuracy on the comprehension questions in the sentence-processing experiments, $r(789) = .097$, $p = .006$. However, the comprehension questions in those experiments were not designed to evaluate depth of understanding, but to provide participants with a task goal in experiments designed to study eye movements during sentence reading. Average scores on the comprehension questions were high (90.93 %) and also had a very small correlation with gaze durations on words while reading the sentences, $r(789) = –.096$, $p = .007$.

Factor analysis

Exploratory factor analyses, using the program IRTPRO (Cai, Thissen, & du Toit, 2011), were performed on the responses to author names as a way of assessing the dimensionality of the ART. Initial analyses of all 65 author names revealed a two-factor structure, but examination showed that the small second factor loaded mostly on difficult items with very few responses (mean selection rate of 6.5 %). The items that loaded highest on this second factor had low discrimination (described below in the IRT section) and correlated with selection of errors at .44, which is higher than the correlation between all names and errors, $r(1012) = .22$, $p < .001$. We believe that this second factor measured the propensity to guess, because few participants recognized these author names. Accordingly, 15 of the items with loadings greater than .4 on the second factor were removed, and factor analysis was used again on the remaining 50 items.

The second analysis showed that a two-factor structure gave a better fit than did a single-factor model, $G^2(49, N = 1{,}012) = 390.89$, $p < .001$. Table 4 shows the factor loadings after an oblique CF quartimax rotation, which suggests a correlation of .55 between the factors. Factor 1 included Saul Bellow, Thomas Pynchon, Bernard Malamud, Virginia Woolf, Gabriel Garcia Marquez, and Ernest Hemingway.

Table 2 Proportions of author names correct and item response theory (IRT)-estimated parameters: *a* parameters (discrimination), *b* parameters (difficulty), and standard errors

Serial Position	Author Name	Percent Selected	*a* Parameter	*b* Parameter
1	Ernest Hemingway	92.2	1.97 (0.24)	–1.88 (0.15)
2	F. Scott Fitzgerald	89.8	1.49 (0.18)	–1.94 (0.17)
3	Stephen King	83.7	1.08 (0.13)	–1.83 (0.18)
4	T. S. Elliot [a]	80.4	1.51 (0.15)	–1.29 (0.11)
5	J. R. R. Tolkien	77.6	1.87 (0.18)	–1.03 (0.08)
6	George Orwell	72.9	1.55 (0.14)	–0.91 (0.08)
7	Maya Angelou	66.3	0.72 (0.09)	–1.05 (0.15)
8	William Faulkner	62.8	1.55 (0.13)	–0.49 (0.06)
9	E. B. White	55.3	0.88 (0.09)	–0.29 (0.09)
10	Harper Lee	53.0	0.95 (0.09)	–0.16 (0.08)
11	Tom Clancy	51.9	1.12 (0.10)	–0.10 (0.07)
12	J. D. Salinger	47.2	1.53 (0.13)	0.09 (0.06)
13	James Patterson	47.1	0.66 (0.08)	0.18 (0.10)
14	Virginia Woolf	46.7	1.20 (0.11)	0.13 (0.06)
15	John Grisham	43.2	1.40 (0.12)	0.25 (0.06)
16	Ray Bradbury	37.7	1.20 (0.10)	0.52 (0.07)
17	Thomas Wolfe	37.3	0.66 (0.08)	0.86 (0.14)
18	Jack London	36.5	1.12 (0.10)	0.61 (0.08)
19	Toni Morrison	35.1	0.59 (0.08)	1.12 (0.17)
20	Ayn Rand	35.1	1.47 (0.12)	0.57 (0.06)
21	John Irving*	32.5	0.74 (0.10)	1.09 (0.17)
22	James Joyce	30.0	1.39 (0.12)	0.81 (0.07)
23	Kurt Vonnegut	27.1	1.97 (0.17)	0.79 (0.06)
24	Samuel Beckett	26.4	0.75 (0.09)	1.52 (0.18)
25	Margaret Atwood	22.4	0.99 (0.10)	1.48 (0.14)
26	Danielle Steel	20.7	0.94 (0.10)	1.66 (0.16)
27	Ralph Ellison	19.7	0.75 (0.09)	2.08 (0.24)
28	Gabriel Garcia Marquez	18.3	1.17 (0.12)	1.59 (0.13)
29	Alice Walker*	16.7	0.77 (0.12)	2.32 (0.33)
30	Isabel Allende	14.4	0.98 (0.11)	2.12 (0.20)
31	Isaac Asimov	13.5	1.56 (0.15)	1.64 (0.11)
32	T. C. Boyle*	13.1	0.30 (0.11)	6.44 (2.44)
33	Vladimir Nabokov	12.2	1.36 (0.14)	1.89 (0.14)
34	Joyce Carol Oates	11.9	1.21 (0.13)	2.06 (0.17)
35	Margaret Mitchell*	11.3	0.86 (0.13)	2.73 (0.39)
36	Clive Cussler	10.9	1.07 (0.13)	2.35 (0.22)
37	Robert Ludlum	8.6	1.22 (0.14)	2.4 (0.21)
38	Salman Rushdie	7.4	1.71 (0.19)	2.09 (0.14)
39	Willa Cather	7.0	1.53 (0.17)	2.28 (0.17)
40	Nora Ephron	6.6	1.26 (0.16)	2.61 (0.24)
41	Jackie Collins*	6.1	0.58 (0.17)	4.93 (1.35)
42	Sue Grafton	5.1	1.10 (0.16)	3.13 (0.35)
43	Kazuo Ishiguro	4.8	1.18 (0.17)	3.05 (0.32)
44	Anne McCaffrey	4.8	1.13 (0.16)	3.15 (0.35)
45	Paul Theroux*	4.8	0.74 (0.20)	4.36 (1.08)
46	Judith Krantz	4.7	1.05 (0.16)	3.33 (0.40)
47	Thomas Pynchon	3.6	2.52 (0.39)	2.28 (0.14)
48	James Michener	3.5	1.18 (0.19)	3.38 (0.40)

Table 2 (continued)

Serial Position	Author Name	Percent Selected	*a* Parameter	*b* Parameter
49	Ann Beattie*	2.6	0.59 (0.26)	6.49 (2.72)
50	Michael Ondaatje	2.6	1.14 (0.20)	3.75 (0.51)
51	Nelson Demille [b]	2.4	1.12 (0.21)	3.87 (0.56)
52	Umberto Eco	2.3	1.57 (0.25)	3.15 (0.31)
53	Raymond Chandler	2.2	1.45 (0.24)	3.34 (0.37)
54	Dick Francis*	2.0	0.30 (0.30)	13.09 (12.75)
55	Sidney Sheldon*	1.9	0.57 (0.33)	7.18 (3.87)
56	Saul Bellow	1.8	2.39 (0.39)	2.77 (0.20)
57	James Clavell*	1.8	0.77 (0.35)	5.62 (2.28)
58	Jonathan Kellerman*	1.8	1.03 (0.29)	4.42 (1.17)
59	Wally Lamb*	1.6	1.36 (0.30)	3.74 (0.63)
60	Jane Smiley	1.4	0.30 (0.27)	14.34 (12.76)
61	Jean M. Auel*	1.0	0.29 (0.36)	16.00 (20.28)
62	Brian Herbert*	0.6	0.15 (0.54)	34.44 (124.95)
63	Tony Hillerman*	0.4	1.00 (0.70)	6.06 (3.98)
64	Herman Wouk	0.4	1.99 (0.49)	3.91 (0.52)
65	Bernard Malamud	0.3	1.80 (0.53)	4.28 (0.74)

* Items excluded from 50-item IRT analyses, but 65-item IRT parameters are presented for comparison. [a] Minor misspelling in ART version; should read "T. S. Eliot." [b] Minor misspelling in ART version, should read "Nelson DeMille."

Books by these authors are regarded as having high literary value, and it seems plausible that many of the participants were exposed to their works through school curricula. Factor 2, with fewer authors with high loadings, included Herman Wouk, Robert Ludlum, Clive Cussler, Tom Clancy, Nelson DeMille, and J. R. R. Tolkien. Clancy, Ludlum, Cussler, and DeMille are all known for their popular thrillers. Wouk's books include *The Caine Mutiny* and *The Winds of War*, Ludlum wrote the *Bourne* series, and Tolkien wrote *The Lord of the Rings* trilogy; all of these books were adapted into popular movies. Some authors who loaded on Factor 2, like Anne McCaffrey, did not write books that were adapted into movies, but their books (e.g., the *Dragonriders of Pern* series) were likely to have been encountered outside of the classroom. The results of the factor analysis suggest that the list of authors on this ART has the potential to measure individuals' knowledge of popular and literary authors separately. This possibility should be viewed with some caution, since it is based on subjective, post-hoc classification of authors into the popular and literary categories. In addition, some authors do not fit this subjective classification. For example, Danielle Steel (*Season of Passion*) loads on the first (literary) factor, but not on the second (popular) factor.

Although our item factor analysis provides a potential basis for treating this ART as having distinguishable factors (literary vs. popular), our subsequent analyses treated the test as unidimensional because the factors were correlated and not widely divergent. All of the results reported below for the 50-item ART are very similar to those found with the full 65-item ART.

Table 3 Percentages of participants who selected different numbers of foils (false-alarm errors)

Errors	%
0	66.6
1	18.5
2	6.3
3	3.4
4	1.6
5	1.2
6–19	2.5

$N = 1{,}012$

IRT, ART, and item selection

IRTPRO (Cai et al., 2011) was used to estimate the parameters of a two-parameter logistic (2PL) model for the pooled sample of 50 items. The 2PL model resulted in excellent fit, M_2(df = 1175) = 2,244.94, $p < .001$, RMSEA = .03.[2] Both simpler (1PL) and more complex (3PL) models were also assessed, but these showed poorer fits than did the 2PL model and are not considered further. The 2PL model (shown in the equation

[2] The 2PL model with 65 items also fit well: M_2(df = 2015) = 4,319.63, $p < .001$, RMSEA = .03. Thus, IRT analyses for the 15 items that were excluded are reported in Table 2.

Table 4 Factor analysis: 50-item loadings from a two-factor exploratory factor analysis with oblique rotation and standard errors

	Name	Factor 1	Factor 2
1	Saul Bellow	**–0.84** (1.13)	0.02 (2.11)
2	Thomas Pynchon	**–0.75** (1.40)	–0.13 (2.12)
3	Bernard Malamud	**–0.73** (3.34)	–0.01 (5.52)
4	Willa Cather	**–0.71** (0.91)	0.04 (1.77)
5	Virginia Woolf	**–0.71** (0.62)	0.15 (1.62)
6	Gabriel Garcia Marquez	**–0.71** (0.55)	0.18 (1.58)
7	Kurt Vonnegut	**–0.69** (1.28)	–0.12 (1.93)
8	J. D. Salinger	**–0.69** (0.97)	0 (1.77)
9	Ernest Hemingway	**–0.68** (1.29)	–0.13 (1.93)
10	F. Scott Fitzgerald	**–0.68** (0.95)	0 (1.74)
11	Vladimir Nabokov	**–0.67** (0.82)	0.04 (1.67)
12	Salman Rushdie	**–0.66** (1.17)	–0.10 (1.82)
13	Kazuo Ishiguro	**–0.65** (0.66)	0.10 (1.53)
14	William Faulkner	**–0.64** (1.08)	–0.07 (1.74)
15	Margaret Atwood	**–0.63** (0.52)	0.15 (1.40)
16	Ayn Rand	**–0.62** (1.05)	–0.07 (1.69)
17	James Joyce	**–0.61** (1.00)	–0.06 (1.64)
18	Joyce Carol Oates	**–0.61** (0.78)	0.03 (1.52)
19	Nora Ephron	**–0.59** (0.88)	–0.02 (1.54)
20	Michael Ondaatje	**–0.55** (0.89)	–0.02 (1.50)
21	Toni Morrison	**–0.55** (0.13)	0.27 (1.03)
22	T. S. Elliot	**–0.54** (1.25)	–0.20 (1.65)
23	Harper Lee	**–0.52** (0.69)	0.02 (1.31)
24	George Orwell	**–0.50** (1.35)	–0.25 (1.64)
25	Isabel Allende	**–0.50** (0.74)	–0.05 (1.33)
26	Ralph Ellison	**–0.49** (0.44)	0.10 (1.13)
27	Danielle Steel	**–0.48** (0.74)	–0.02 (1.28)
28	Maya Angelou	**–0.47** (0.44)	0.08 (1.08)
29	Ray Bradbury	**–0.46** (1.09)	–0.17 (1.43)
30	E. B. White	**–0.46** (0.71)	–0.03 (1.21)
31	Umberto Eco	**–0.45** (1.50)	–0.33 (1.67)
32	Isaac Asimov	**–0.40** (1.56)	**–0.39** (1.58)
33	Raymond Chandler	**–0.39** (1.47)	–0.36 (1.53)
34	Samuel Beckett	**–0.39** (0.66)	–0.04 (1.05)
35	Stephen King	**–0.37** (1.13)	–0.24 (1.28)
36	James Patterson	**–0.27** (0.73)	–0.14 (0.89)
37	Thomas Wolfe	**–0.25** (0.76)	–0.16 (0.87)
38	Jane Smiley	**–0.14** (0.65)	–0.05 (0.69)
39	Herman Wouk	0.05 (2.54)	**–0.99** (1.43)
40	Robert Ludlum	0.05 (2.14)	**–0.86** (1.08)
41	Clive Cussler	0.05 (1.96)	**–0.79** (0.99)
42	Tom Clancy	–0.03 (1.95)	**–0.75** (1.13)
43	Nelson Demille	–0.01 (1.80)	**–0.70** (1.01)
44	J. R. R. Tolkien	–0.36 (1.90)	**–0.55** (1.68)
45	Jack London	–0.19 (1.58)	**–0.52** (1.21)
46	James Michener	–0.18 (1.61)	**–0.52** (1.21)
47	John Grisham	–0.32 (1.61)	**–0.46** (1.46)
48	Anne McCaffrey	–0.23 (1.48)	**–0.45** (1.24)
49	Sue Grafton	–0.22 (1.45)	**–0.44** (1.20)
50	Judith Krantz	–0.21 (1.39)	**–0.43** (1.15)

Factor intercorrelation of $r = .55$

below) provides information about item discrimination (a) and item difficulty (b) in relation to the underlying ability (θ).

$$P(\text{correct selection}) = \frac{1}{1+\exp[-a(\theta-b)]}$$

The discrimination parameter (a) is the slope of an item's logistic function at its point of inflection. It provides information about how well each item discriminated between respondents of low ability and those of high ability. Items with discrimination parameters close to 0 are unlikely to contribute much to the test. The b parameter, item difficulty, represents the level of ability that an individual had to possess to have a 50 % chance of correctly responding to an item. Difficult items have high b parameters. The underlying ability (θ) is the participant's ability, which is estimated by the test. For the ART, this underlying ability is most directly characterized as the ability to recognize authors' names, with that ability hypothesized to depend on print exposure (Stanovich & West, 1989). The parameters are reported in Table 2.

Using the responses to items, IRT estimated the parameters and the *item characteristic curves* (ICCs) for each item. The ICCs, also called *trace lines*, show the probabilities of correctly responding to a specific item at any level of estimated latent ability. The left panel of Fig. 1 shows the ICCs for four items that were effective, in the sense that they have high discrimination as indicated by a. The four items also progress in difficulty from Hemingway to Tolkien to Vonnegut to Bellow, as indicated by their b values. A very easy item, like Hemingway, provides information that is useful in distinguishing among participants at the lower range of abilities, but little information about participants at higher levels of ability, because almost all of them will correctly select him as an author. In contrast, a very difficult item, like Bellow, provides information that is useful in distinguishing among participants at high levels of ability, but little information about participants at lower levels of ability, since almost none of them will correctly select him as an author. The right panel of Fig. 1 shows the ICCs for four relatively ineffective items (which have low discrimination).

Item information functions are representations of the amount of information that an item provides along a range of abilities, derived from the ICCs and the population distribution. These functions tend to look bell-shaped, with tall functions contributing more information. Figure 2 shows the *test information function*, which sums the information

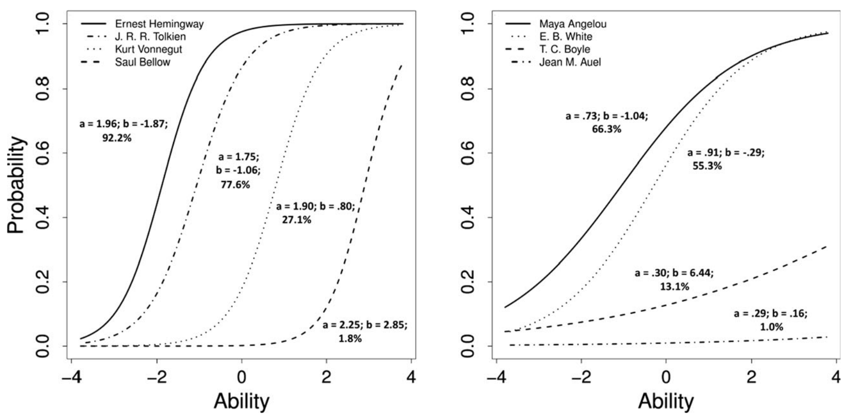

Fig. 1 Item characteristic curves of effective author recognition test (ART) items (left) and items with low discrimination (right) using the 65-item ART parameters. Note that the 50-item ART parameters were very similar, but T. C Boyle and Jean M. Auel were omitted from the shorter test

functions for the 50-item test. It indicates that the test items provide high amounts of information about high scorers, but low information about low scorers, resulting in relatively low precision when estimating scores for participants low in ability. This means that this ART has an imbalance, with too few easy items such as Tolkien, and too many difficult items such as Bellow. Using the full 65-item ART does not correct this imbalance, since most of the 15 additional authors were very difficult.

The 2PL IRT model was fit to the author names, but not to the foils. Inclusion of foils led to difficulty in interpreting the model, because most of the foils had very low rates of selection and because there was a moderate correlation between the number of ART names selected and the number of foils selected, even after removing the 15 author names that were most highly correlated with guessing, $r(1012) = .16, p < .001$. This positive relationship between number of hits and number of false alarms suggests that participants varied in the strictness of their criteria for making an "author" judgment. That is, some participants strictly obeyed the instruction not to guess, and therefore needed to be absolutely certain that they knew an author before selecting the name, whereas other participants were more lenient, and therefore made errors due to incorrect guesses. However, because the majority of participants (66.6 %; see Table 3) made no false-alarm errors, there was no ART-internal method for determining whether the standard ART scoring method (number of authors selected minus number of foils selected) would provide the optimal treatment of criterion differences. In the analyses to be reported below, we used the gaze duration data as an external criterion for how hits and false alarms should be scored in the ART.

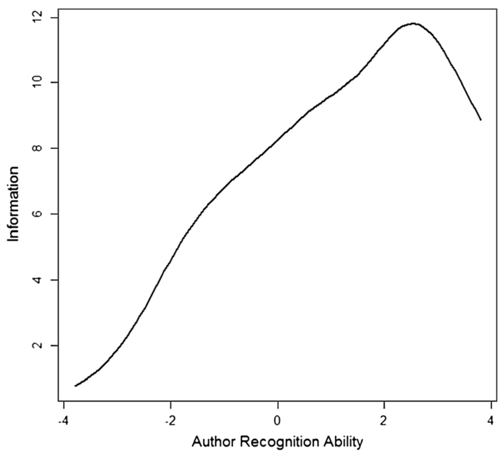

Fig. 2 Test information function for the 50-item ART

Eyetracking measures

The present analyses focused on gaze duration, which is related robustly to measures of word difficulty, such as word frequency, and is commonly treated as the best eyetracking measure of word recognition (Rayner, 1998). Gordon et al. (2014) showed that ART is related to gaze duration and provide detailed analyses of the relation of ART to other eye movement measures of both early lexical processing and later processes of sentence comprehension. Those more detailed analyses were performed using data from a subset of the present data in which participants read a common set of sentences. In the present study, variation across the sentences and words read by different participants was addressed through the inclusion of word frequency (and the associated word length variation) in statistical models of gaze duration. Figure 3 illustrates the relationship of gaze duration to ART performance by dividing participants into groups on the basis of their knowledge of two authors (J. R. R. Tolkien and Kurt Vonnegut), who were chosen for this purpose because they have high discrimination (see Fig. 1) and because their difficulty splits the participant population into three large groups. The effect of word frequency on gaze duration is illustrated by placing the words into 20 bins that are equally sized on log word frequency. Mean gaze duration declined significantly

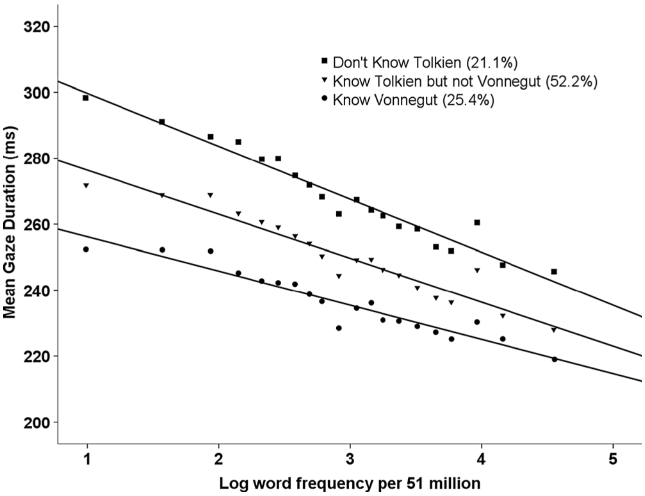

Fig. 3 Relationship between word frequency and mean gaze duration as a function of author knowledge: Words were grouped by log frequency into 20 equally sized bins. Consistent with the item characteristic curves shown in Fig. 1, only 1.3 % of participants selected Kurt Vonnegut but not J. R. R. Tolkien

with increasing author knowledge (269 ms for those knowing neither author, 250 ms for those knowing Tolkien but not Vonnegut, and 235 ms for those knowing both), $F(2, 776) = 34.8$, $p < .001$. The magnitude of the differences in mean gaze durations between the three groups should be interpreted with the understanding that 200 ms is generally considered the minimum duration for a voluntarily controlled fixation (Rayner, 1998). Figure 3 also shows a large effect of word frequency (and the associated variation in word length) on gaze durations for all participant groups, and that the magnitude of the increase in gaze duration for low-frequency words increases as author knowledge decreases.

Scoring rule methods

Table 5 shows the correlation between various ways of scoring the ART and three reading-time measures. The first measure is mean gaze duration, and the second two are the intercept and slope parameters from regression analyses for individual participants in which log word frequency was used as a predictor of gaze duration. For all of the scoring rules, higher ART scores were associated with faster reading times (shown by both means and intercepts) and with a smaller effect of word frequency on gaze duration. The results of these continuous analyses confirm the patterns seen for the partition of participants into the three groups shown in Fig. 3.

If gaze duration, which reflects the efficiency of word recognition, is taken as an indicator of language skill, then alternative scoring rules for the ART can be evaluated by how well they predict gaze durations. Here this was done to evaluate classical test theory (summed scores) and IRT as ways of scoring hits (selection of authors), and further, to evaluate different penalties for false alarms (selection of foils). IRTPRO was used to determine score estimates for each of the 1,012 participants, using the recommended method for estimating IRT scores, *expected a posteriori* (EAP). The means and standard deviations of these new scores were matched with those of the summed scores of 50 ART names without a foil penalty (the "50 ART name score" in Table 1). After estimating individual scores, error penalties of varying degrees were applied, and the resulting scores were correlated with the measures of participants' gaze durations (789 participants in total). Table 5 shows the results for no penalty, losing 1 for every false alarm, and losing 2 for every false alarm. More severe penalties led to declines in the correlations.

As Table 5 shows, the combination of IRT-based EAP scores with a two-author penalty for every foil selected had a slightly higher correlation (–.39) with mean gaze duration than did the standard score (summed score with a one-author penalty; –.38), Z_H = (786) 2.01, p = .044, by Steiger's Z test (Hoerger, 2013). EAP scores with a two-author penalty also showed stronger relations than the standard score with intercepts, -.37 (EAP; intercept; -2 penalty) versus –.35, Z_H = (786) 2.23, p = .026, and with frequency slopes, .24 versus .22, Z_H = (786) 2.02, p = .043. The differences between the alternative scoring methods are small and unlikely to have much impact on most studies. However, they are significant, given the present large sample size, and are consistent across measures. The correlation increases suggest that taking into account the discriminative ability of individual items may lead to estimation of more accurate ART scores.

Frequency and item difficulty

Stanovich and West (1989) reasoned that the ART measures print exposure because the likelihood of encountering authors' names increases with their amount of reading. As we discussed above, support for this important characterization has come from a variety of sources (e.g., diary studies), but any evidence about this relationship is necessarily indirect. Here we addressed this issue by examining the relationship between author difficulty and the frequency with which the author's name appears in samples of written English. If ART performance relates to amount of reading, then the difficulty of an ART author should decrease as a function of the frequency with which the author's name appears in print. This decrease should occur because the amount of reading required in order to encounter the author's name enough times for it to become familiar would decrease as a function of the frequency of the author's name.

IRT measures item difficulty through the b parameter, which is defined as $b = –c/a$.[3] Thus, item difficulty can be artificially inflated when discrimination (the a parameter) is low. One author (Jane Smiley) was excluded from analyses of

[3] The c parameter is a computationally simpler estimate of difficulty.

Table 5 Correlation between ART and gaze duration (GZD) measures, as a function of ART scoring method, using the 50-item ART

	Summed Score			IRT (EAP) Score		
	Mean GZD	GZD Intercept	GZD Frequency Slope	Mean GZD	GZD Intercept	GZD Frequency Slope
No Penalty	–.35	–.32	.20	–.36	–.33	.21
Lose 1 per false alarm	–.38	–.35	.22	–.39	–.36	.23
Lose 2 per false alarm	–.39	–.36	.24	–.39	–.37	.24

$N = 789$. The first measure, mean GZD, is the average gaze duration. The second two measures, GZD intercept and GZD frequency slope, are parameters from regression analyses performed on the individual participants using log word frequency as a predictor. All correlations in the table were significant at the $p < .001$ level. Scores from the 65-item test showed the same pattern, with slightly decreased correlations.

the relationship between item difficulty and author name frequency because its very low *a* parameter resulted in an inflated *b* parameter.

Two sources of data were used to estimate author name frequency. The first estimate was made using the Google Terabyte *N*-Gram Corpus (Brants & Franz, 2006), a sample of nearly one trillion words and character strings from English-language Web sites that Google collected in January 2006 and tabulated into *n*-grams. For authors who were listed in the ART with only a first and last name (e.g., Danielle Steel, Umberto Eco), name frequency was assessed using the two-gram database, whereas for authors listed with a middle name and/or initials (e.g., F. Scott Fitzgerald, Joyce Carol Oates), name frequency was assessed using the three-gram database. Because some names were likely to be referred to in a variety of ways, name spellings that varied slightly from the name presented in the 50-item ART were included for a few names (e.g., "R. Tolkien" and "JRR Tolkien"; "J. D. Salinger" and "JD Salinger"). Item difficulty (*b* parameter) showed a strong relationship to the log of author name frequency, $r(49) = –.71$, $p < .001$. The scatterplot in Fig. 4 shows this relationship with the full 65-item test, as well as the difficulty and frequency estimates for individual authors. The relationship between author difficulty in the Acheson et al. (2008) data and author name frequency was also assessed by correlating the author name frequency with the logits of the author-selection proportions reported in the article. The resulting correlation ($r = –.71$) was higher than that observed for the present sample ($r = –.65$) when difficulty was assessed using logit-transformed proportions, indicating that estimates of frequency for the data, likely collected around 2006–7, were more accurately measured by the Google Terabyte *N*-Gram Corpus.

The second estimate of author name frequency was made using the Corpus of Contemporary American English (COCA), which contains 450 million words drawn from sources published from 1990 to 2012, including fiction, magazines, newspapers, and academic journals (Davies, 2008). COCA is much smaller than the Google Terabyte *N*-Gram Corpus, but its data sources are better understood, and in addition to providing frequency information in response to a query (such as an author's name), it can also randomly select instances that match the query and show the context in which the query text appears. This context can be used to evaluate whether the name is actually being used in reference to the author. COCA was queried with author names as they were listed in the ART, and also with the variations on those names described for the Google Terabyte estimates. In order to estimate author frequency, we input the author names listed in the ART for most authors, and input different name spellings for a few authors (e.g., "F. Scott Fitzgerald" and "Francis Scott Fitzgerald"). In addition, COCA was queried using each author's last name but excluding instances in which it was preceded by the first name or initials. The surrounding contexts were examined for 20 randomly selected matches to each

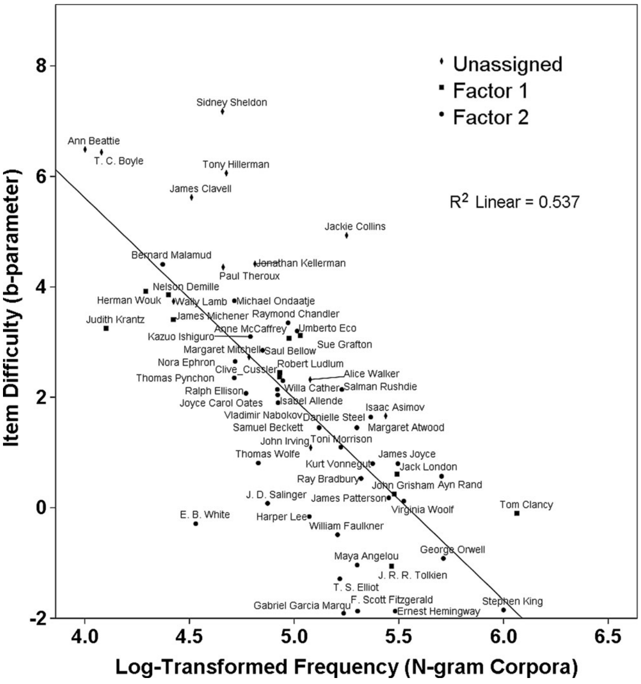

Fig. 4 Relationship between item difficulty and log frequency of the author names: Item difficulty of each author name (*b* parameter) and $\log_{10}$-transformed frequency of the name, $N = 61$, based on the 65-item version of the test in order to show items not assigned a factor. Brian Herbert, Dick Francis, Jean M. Auel, and Jane Smiley were not included due to *b* parameters greater than 10

author's full and last names, in order to determine the proportions of cases in which these different ways of naming were used for the author. The proportions and counts were arithmetically combined into a composite frequency estimate based on the full-name and last-name-only author references. We found a significant negative correlation between log frequency and item difficulty, $r(49) = -.61$, $p < .001$, which is similar to yet slightly smaller than the correlation obtained using the Google Terabyte *N*-Gram Corpus.[4] COCA may show a decreased correlation for a variety of reasons, including the smaller size of the database, the range of time periods, or the inclusion of spoken sources.

Although the 50-item test was better for scoring purposes, the 65-item test appeared to relate more strongly to frequency [see Fig. 4; relationship between log frequency and item difficulty in Google Terabyte *N*-Gram Corpus, $r(61) = -.73$, $p < .001$, and in COCA, $r(61) = -.71$, $p < .001$]. This is likely because the 15 items that we removed had low selection rates and low *a* parameters, but addressed a greater level of difficulty than was seen in the shortened test. Our final sample size of 49, while still high enough to gauge the relationship between author name frequency and difficulty, did not capture all potential variance in difficulty.

Although the initial factor analysis pointed to a two-factor interpretation of this version of the ART, our subsequent analyses treated it as unidimensional on the assumption that both factors measure print exposure. As a test of this idea, author difficulty was correlated with author frequency separately within each factor. The 36 items that loaded on Factor 1 (literary) were assigned to that factor, and the 12 items that loaded highest on Factor 2 (popular) were assigned to it. Isaac Asimov loaded equally on the two factors, and accordingly was not assigned to a factor, whereas Jane Smiley was excluded due to an inflated *b* parameter (see Fig. 4). Author difficulty correlated with author frequency to similar degrees for both factors [first factor: Google *N*-Gram, $r(36) = -.65$, $p < .001$; COCA, $r(36) = -.44$, $p = .001$; second factor: Google *N*-Gram, $r(12) = -.87$, $p < .001$; COCA, $r(12) = -.87$, $p = .005$]. The differences between these correlations were not significant in the *N*-Gram analyses, a pattern that is consistent with the conceptualization of the ART as a unidimensional measure of an underlying ability based on print exposure. However, the difference was significant for the COCA analyses, $z = -2.29$, $p = .011$. We found some evidence that the second factor related more highly to word frequency than did the first, although utilizing only 12 names for the second factor made this interpretation unreliable.

Frequency change and author recognition The strong relationship between item difficulty and author name frequency suggests that changes in the pattern of item difficulty for participants taking the test in different time periods might be due to changes in the relative frequencies of author name use across those time periods. As we noted above, for the 17 authors that appeared in the original ART evaluations, the correlation of selections in the present North Carolina study and the Acheson et al. (2008) Wisconsin study was very high, $r(17) = .87$, whereas it was substantially lower with the selections reported earlier by Stanovich and colleagues, $r(17) = .27$. On the basis of publication dates, it is reasonable to infer that the North Carolina and Wisconsin data were collected about 5–7 years apart, whereas the original ART data were collected some 25 years earlier. Google Ngram Viewer, an online word/phrase frequency-graphing tool based on words from millions of books published between 1500 and 2008 (Michel et al., 2011), was used to evaluate whether changes in the frequencies of authors' names predicted changes in author difficulty between Stanovich's earlier data and the more recent data. Use of the authors' names was estimated using Google Ngram Viewer (Google Books, 2013) data for books published from 1979 to 1988 for the earlier period, and from 1999 to 2008 for the recent period. The change in frequency for each of the 17 author names was calculated as the difference between its log proportions in the recent and the earlier corpora.[5] Changes in author difficulty were calculated as the difference between their logit-transformed selection proportions in the recent (North Carolina and Wisconsin) and earlier (Stanovich) studies. The relationship between changes in author difficulty and frequency was very strong, both for the North Carolina data, $r(17) = .73$, $p = .001$, and for the Wisconsin data, $r(17) = .81$, $p < .001$. These relationships are even stronger when the analysis is restricted to the 11 authors who were retained in our 50-item ART, as is shown in the left panel of Fig. 5 for the North Carolina data and in the right panel of that figure for the Wisconsin data (*r*s = .86 and .94, respectively).

Examination of Fig. 5 suggests possible explanations for some of the authors whose changes in difficulty deviated from their changes in frequency. In both samples, Tolkien was selected more often than expected, possibly because the release of popular movies based on *The Lord of the Rings* and *The Hobbit* increased his exposure in media not captured in the Google Ngram Viewer. In addition, the North Carolina sample shows higher-than-expected selection of Thomas Wolfe, a result that can be plausibly attributed to his being a

[4] Logit-transformed mean selection rates of the author names could be used instead of the IRT-produced *b* parameters in the above analyses. The main benefits of using IRT analyses are the identification of effective author names and the removal of items with inflated *b* parameters. Utilizing IRT also appears to better represent the relationship between difficulty and author frequency for items with low selection rates. Correlations of the logit-transformed proportions with the same frequency estimates were still significant with both the Google *N*-gram data set [$r(49) = .65$, $p < .001$] and the COCA data set [$r(49) = .54$, $p < .001$].

[5] All author names were input as they were spelled in the Stanovich study, except J. R. R. Tolkien (input as "Tolkien").

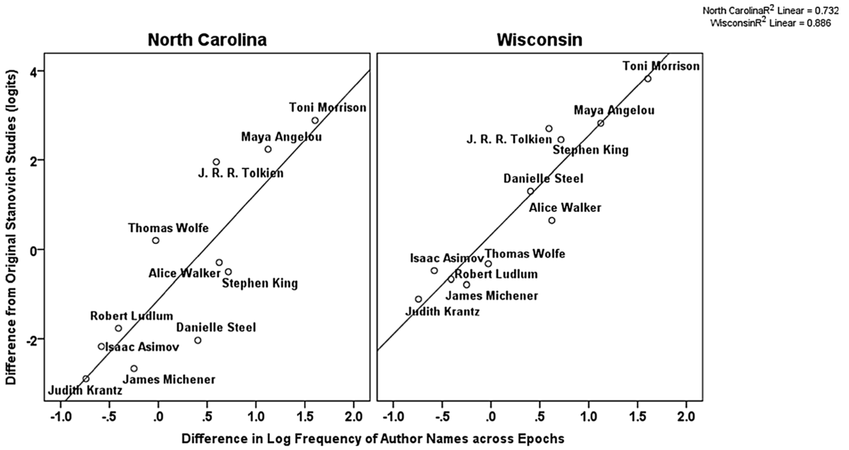

Fig. 5 Differences in author log frequencies from the Google Ngram Viewer (1999–2008 and 1979–1988) by difference in the logit-transformed mean selection rates per author. The panels show differences between the present study and the Stanovich studies (Stanovich & Cunningham, 1992; Stanovich & West, 1989; left) and differences between Acheson et al. (2008) and the Stanovich studies (right)

North Carolina native whose first successful novel was largely set in North Carolina. A similar pattern is seen for Maya Angelou, who lived and taught in North Carolina throughout the lifetimes of the study participants. For other deviations from the trend line (e.g., the lower-than-expected knowledge of Danielle Steel), no obvious explanation comes to mind.

The consistency and predictability of the differences in author difficulty from the early (Stanovich & Cunningham, 1992; Stanovich & West, 1989) to the recent data sets raises the question of how difficulty is related in the recent data sets. As we noted above, there is a high correlation [$r(65) = .88$] between author selections in the present data and the Acheson et al. (2008) data. Assessment of whether the differences in author difficulty between the two data sets are due to changes over time in the frequencies of the authors' names is made difficult by the relatively brief period of time between the two studies and because the Google Ngram Viewer used to assess changes in author name frequency only extends through 2008, which means that frequency data were not available for most of the time between the two studies. Although the relative difficulties of author items were highly similar, the overall selection rates of author names were higher in the Acheson et al. study (36 %) than in the present study (24 %). The reasons for this substantial difference are not clear. It seems unlikely that the difference was due to differing admissions standards, since standardized test scores tend to be slightly higher at the University of North Carolina than at the University of Wisconsin.[6] One possibility is that differences in participant sampling methods caused the difference. The present study tested students from Introductory Psychology classes, which primarily enroll first- and second-year students, whereas the Acheson et al. study recruited paid participants from the larger university community, which may have resulted in a greater proportion of older students who had progressed further in their educations. Finally, it is possible that the test has become harder because some of the authors who were considered sufficiently prominent for inclusion when the test was constructed have become less prominent since the test was constructed. Evidence in support of this possibility comes from a decline over time in ART scores within the present study. On average, participants tested in the first half of the study (fall 2010 through spring 2012) had higher scores than did participants tested in the second half of the study (fall 2012 through spring 2014) (15.3 vs. 14.2), $t(1010) = 2.43$, $p = .015$.

Discussion

In the research reported here, we used item response theory to investigate the psychometric properties of the author recognition test and to investigate two substantive issues: the relationship between performance on the ART and speed of word recognition during reading, and how item (author) difficulty on the ART is related to the frequency with which the author's name appears in print. Analysis of the effectiveness of individual items (authors) indicated that, at least in studies targeting young adults, 15 authors should be eliminated from the 65-author ART developed by Acheson et al. (2008) because of their correlation with guessing. We found some evidence that the resulting 50-item ART should be conceptualized as a two-dimensional model, with intercorrelated factors that could be interpreted as distinguishing popular and literary authors. There were highly significant relations of performance on the ART with gaze durations on words and with the effect of word frequency on gaze durations. The

[6] In the years preceding collection of the ART data, the typical interquartile ranges for SAT Verbal and ACT English scores were 590–700 and 26–33 at the University of North Carolina (2011), as compared to 550–670 and 25–31 at the University of Wisconsin–Madison (2008).

strength of these relationships was slightly greater when ART performance was assessed using an IRT measure of author knowledge in combination with a two-point penalty for incorrectly selecting nonauthors, as compared to the summed score assessment that has been standard in applications of the ART. Variation in item difficulty was strongly related to variation in the frequencies of authors' names in a large text sample, and changes in item difficulty, shown by differences between early and more recent ART data, were strongly related to differences in the frequencies of authors' names in the decades preceding data collection. These strong relations between item difficulty and frequency provide novel evidence in support of the argument by Stanovich and West (1989) that ART performance reflects time spent reading.

ART, print exposure, and word recognition

Stanovich and West (1989) developed the ART as a measure of print exposure—time spent reading—on the rationale that knowledge of authors' names was likely acquired through reading, so that individuals who read more were more likely to have encountered authors' names and to remember them. A number of findings (e.g., the relation of ART performance to reports of time spent reading in diary studies) provide support for this thesis (see the introduction). The present study tested this characterization of the ART by determining the relation between item difficulty and author name frequency, according to the rationale that if author knowledge is derived from reading, then author difficulty should be inversely related to the print frequency of the author's name. This is because more reading would be required to encounter authors whose names appear infrequently than to encounter those whose names appear frequently. As is shown in Fig. 4, item difficulty decreased as author name frequency increased. In addition, as is shown in Fig. 5, comparisons between these recent data sets and ones collected some 25–30 years earlier (Stanovich & Cunningham, 1992; Stanovich & West, 1989) showed that changes in the difficulty of psychometrically valid author items were strongly related to changes in the log frequencies with which the authors' names appeared in books published in the decades prior to data collection (1979–1988 and 1999–2008), as measured using the Google Ngram Viewer (Michel et al., 2011). Estimates of differences over time in the frequencies of authors' names accounted for 73.2 % of the variance in the present study, and an impressive 88.6 % with the Acheson et al. (2008) study.

The very strong relations observed here between author difficulty and author name frequency (see Fig. 4; frequency accounts for 53.7 % of the variance in author difficulty) supports the ideas that although the ART is a direct test of a very specific kind of knowledge, it has value as a reading test because it is an indirect test of how much practice with reading people have had. The eyetracking results in the present study show that ART scores predict gaze durations on words during reading as well as the effect of word frequency on gaze durations during reading (see Fig. 3 and Table 5). It is widely recognized that word recognition is an essential step in reading comprehension and that efficient word recognition processes free up cognitive resources for higher levels of language processing (Perfetti, 2007). The present findings are consistent with earlier ones showing that the ART predicts a variety of lexical skills in isolated word recognition tasks (Chateau & Jared, 2000; Sears et al., 2006; Stanovich & West, 1989), but it extends those findings to natural reading.

Psychometrics of the ART

The IRT and factor analysis results reported here provide important information about the Acheson et al. (2008) ART and identify three issues that should be addressed in future versions of the test. First, as is shown in Fig. 2, the author items are far more informative about differences at higher than at lower levels of ability; future versions of the test should include a greater number of easy author items, in order to provide more information about lower levels of ability. Second, exploratory factor analysis (Table 4) supported a conceptualization of the ART as a two-dimensional model, but further analysis will be needed to assess the validity of this characterization and to determine whether such a distinction has predictive importance for the ART. The initial development of the ART aimed to choose only authors that participants were likely to have read in leisure time and to avoid authors that were regularly studied in the classroom (West et al., 1993). If the first factor does indeed measure academic or literary reading, and if it does not relate to print exposure as highly as the second factor does, as is suggested by the COCA analyses, then this may indicate that more popular authors should be added to the ART. Finally, this study showed a positive relationship between selecting author names and selecting foils, indicating that adopting a lower criterion for making an author judgment leads to higher scores. Increasing the error penalty for selecting a foil from 1 to 2 strengthened the relationship between ART score and gaze durations during reading. However, we had no way to assess criterion variation in the majority of participants (65 %) who made no foil selections. The ART might be improved by the adoption of methods, such as confidence ratings, that would provide more information about the strictness of participants' criteria.

More generally, this study reinforces the importance of frequently updating the ART, and the results of the factor analysis and IRT are useful in differentiating between effective and ineffective items to retain or remove in future versions of the test. If the ART is to be given to a population that is similar to the undergraduates tested in this study, then the 15 items that are thought to measure the propensity for guessing should be replaced. Furthermore, author names that have high discrimination and that cover a range of difficulties should definitely be included in a revised test, whereas authors with low discrimination are candidates for replacement. The present results provide less specific guidance about how the ART should be revised for a population that differs substantially from the one tested here; the full 65-item ART might provide an appropriate starting point for a more general population that includes older adults. Furthermore, since the ART takes such a short time to administer, it is advisable to err on the side of including too many items rather than too few.

The strong relationships found between author difficulty and the frequency of authors' names (Fig. 4) indicates that corpus frequency is likely to be a very valuable tool for estimating the difficulty of authors under consideration for inclusion in the test. Furthermore, the strong relationships found between changes in author difficulty and changes in the frequencies of authors' names indicates that the use of corpora that focus on particular types of texts and time periods may allow for selection of target items so as to create recognition tests that are appropriate for groups of individuals of different ages or who have different backgrounds and experiences.

Conclusion

The ART takes only three minutes or so to administer and predicts a variety of outcomes related to reading efficiency and future reading gains. The difficulty of author items in the ART is strongly related to the frequency with which their names appear in text, a finding that supports the proposal (Stanovich & West, 1989) that the ART measures print exposure (how much an individual reads). ART scores predict the speed with which college students encode words during natural reading, a finding that supports the view that the efficiency of this essential component of reading comprehension is related to practice at reading. Recognition of particular authors varies substantially over relatively short periods of time. This shows that the test has a great deal of cultural specificity and should be regularly reformulated so that it is appropriate for the group being assessed. However, the selection of author items can be guided by the frequencies of authors' names in different corpora. The present study provides a basis for further test development by showing a major determinant of item difficulty in the ART.

Author Note This research was support by Grant Nos. R01HD060440-06A2 and 3R01HD060440-09S1 to P.C.G. from NICHD. We thank Victor Kuperman for discussion at the early stages of this project and Dave Thissen for advice on the IRT/factor analyses and for comments on a draft of this article. We also thank Wonil Choi, Renske Hoedemaker, Matt Lowder, and Giulia Pancani for discussion of this project and comments on earlier drafts, as well as members of the Language Cognition and Brain group at UNC Chapel Hill for help with data collection and management.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40,* 278–289. doi:10.3758/BRM.40.1.278

Beech, J. R. (2002). Individual differences in mature readers in reading, spelling, and grapheme–phoneme conversion. *Current Psychology, 21,* 121–132.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

Carp, F. M., & Carp, A. (1981). The validity, reliability, and generalizability of diary data. *Experimental Aging Research, 7,* 281–296.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition, 28,* 143–153. doi:10.3758/BF03211582

Cor, M. K., Haertel, E., Krosnick, J. A., & Malhotra, N. (2012). Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey. *Social Science Research, 41,* 1003–1016. doi:10.1016/j.ssresearch.2012.05.007

Cunningham, A. E., & Stanovich, K. E. (1990). Assessing print exposure and orthographic processing skill in children: A quick measure of reading experience. *Journal of Educational Psychology, 82,* 733–740.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990–present. Retrieved April 28, 2014, from http://corpus.byu.edu/coca/

Denckla, M. B., & Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex, 10*(2), 186–202.

Google Books. (2013). Ngram Viewer. Retrieved May 27, 2014, from https://books.google.com/ngrams/

Gordon, P. C., & Hoedemaker, R. S. (2014). *Effective scheduling of looking and talking during rapid automatized naming*. Manuscript under review.

Gordon, P. C., Moore, M., Choi, W., Hoedemaker, R. S., & Lowder, M. (2014). *Individual differences in reading: Separable effects of practice and processing skill*. Manuscript in preparation.

Hoerger, M. (2013). Z_H: An updated version of Steiger's *Z* and web-based calculator for testing the statistical significance of the difference between dependent correlations. Retrieved from www.psychmike.com/dependent_correlations.php

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40,* 431–439. doi:10.3758/BF03208203

Kirby, J. R., Georgiou, G. K., Martinussen, R., Parrila, R., Bowers, P., & Landerl, K. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45,* 341–362. doi:10.1598/RRQ.45.3.4

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading, 31,* 273–284.

Mayer, J. D., Panter, A. T., & Caruso, D. R. (2012). Does personal intelligence exist? Evidence from a new ability-based measure. *Journal of Personality Assessment, 94,* 124–140. doi:10.1080/00223891.2011.646108

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182. doi:10.1126/science.1199644

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137,* 267–296.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11,* 357–383.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422. doi:10.1037/0033-2909.124.3.372

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191–201.

Sears, C. R., Campbell, C. R., & Lupker, S. J. (2006). Is there a neighborhood frequency effect in English? Evidence from reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1040–1062. doi:10.1037/0096-1523.32.4.1040

Sénéchal, M., LeFevre, J.-A., Hudson, E., & Lawson, E. P. (1996). Knowledge of storybooks as a predictor of young children's vocabulary. *Journal of Educational Psychology, 88,* 520–536. doi:10.1037/0022-0663.88.3.520

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literary within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20,* 51–68. doi:10.3758/BF03208254

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24,* 402–433.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81–97. doi:10.1037/1082-989X.1.1.81

University of North Carolina. (2011). Common Data Set 2010–11. Retrieved June 5, 2014, from http://oira.unc.edu/files/2012/03/cds_2010_2011.pdf

University of Wisconsin–Madison. (2008). Common Data Set A: General information (2007–2008). Retrieved June 5, 2014, from http://apir.wisc.edu/publisherssurvey/CDS_2008.pdf

West, R. F., Stanovich, K. E., & Mitchell, H. R. (1993). Reading in the real world and its correlates. *International Reading Association, 28,* 35–50.