Reflecting the Readers of Today: An Update of the American English Author Recognition Task

Ana E. Celmare (acelmare@lst.uni-saarland.de)
Department of Language Science and Technology, Saarland University, Campus C7.2., Saarbrücken, 66123, Germany

Ted J.M. Sanders (t.j.m.sanders@uu.nl)
Institute for Language Sciences, Utrecht University, Trans 10, Utrecht, 3512 JK, Netherlands

Merel C.J. Scholman (m.c.j.scholman@uu.nl)
Institute for Language Sciences, Utrecht University, Trans 10, Utrecht, 3512 JK, Netherlands

Abstract

The Author Recognition Test (ART) is a test that has been widely used for decades to estimate print exposure, due to its reliability and objectivity. In this contribution, we present an updated, balanced, and inclusive version of the English ART. We created this by constructing a contemporary item pool based on recent bestsellers and award-winning authors which balances author genre and literary level. We validate the updated English ART using item-response theory analysis and compare it with a commonly used version for American English, the Acheson et al. (2008) ART. Additionally, we examine how participant-level variables such as age and educational level interact with item-level properties such as genre and literary level to modulate recognition probability. Our results indicate that the updated ART is reliable and the test is comparable with the Acheson version in terms of difficulty and discriminative power. Additionally, reader characteristics and author characteristics systematically modulate print exposure. We discuss the implications of our findings and the suitability of the new version of the ART.

Keywords: author recognition test, print exposure, reading, individual differences

Introduction

Print exposure represents the amount of time a person spends engaging with written language (reading), including reading newspapers, magazines, books, journals, scientific papers, and other printed or digital materials. Research has shown that individuals' levels of print exposure have profound cognitive consequences. For example, higher levels of print exposure tend to correlate with better linguistic skills and better reading comprehension skills (Mol & Bus, 2011; Zufferey & Gygax, 2020; Stoops & Montag, 2023).

Because self-reported reading habits are often unreliable and susceptible to social desirability biases, behavioral proxy measures have been developed to assess print exposure indirectly. One of the most widely used tests is the Author Recognition Test (Stanovich & West, 1989). In the ART, participants are presented with a list of names and asked to indicate which ones they recognize as authors. The list includes both real authors and foils, and performance is scored as the number of correctly recognized authors minus false alarms (that is, foils selected as real authors). The test is based on the assumption that recognizing an author's name does not directly imply that the participant has read the author's work (Moore & Gordon, 2015). Instead, the test measures both primary print knowledge (gained through personal reading) and secondary print knowledge (gained from secondary sources, for example, friends' recommendations, book reviews via blogs or social media, time spent in library/bookstores etc). Both types of print knowledge are positively correlated to better reading comprehension, although primary print knowledge correlates more strongly with reading ability (Martin-Chang & Gould, 2008). Because recognizing an author's name typically requires prior incidental exposure to literary environments (McCarron & Kuperman, 2021), the ART provides a fast and relatively implicit estimate of individuals' cumulative print exposure.

A critical feature of the ART is that performance depends not only on a person's reading experience, but also on the specific authors (and foils) included in the test. Author recognition is dependent on the time and cultural context of when the test is taken: authors rise and fall in prominence, media adaptations change visibility, and reading habits differ across generations. Consequently, the same item set can advantage some age cohorts while disadvantaging others. For example, older participants are less likely to recognize recent authors, whereas younger participants are less likely to recognize authors who were prominent several decades ago (Grolig et al., 2020).

Reading habits and specific literary knowledge are also dependent on the geographical location of the participant sample. For example, French readers are more likely to be knowledgeable of local authors or authors whose work is available in French. As a result, many versions of the ART have been developed for different languages, such as Dutch (Brysbaert et al., 2020), French (Zufferey et al., 2012), German (Grolig et al., 2020; Hug et al., 2024), Turkish (Gedik, 2024) and Chinese (Chen & Fang, 2015).

Some of these studies have also introduced secondary goals to their test development, such as investigating the author- and participant-specific features that lead to differences in recognition rate. In particular, Grolig et al. (2020) explored the variation of print exposure across the lifespan (13 to 77 years old) for a German text, and found that the relation between participant age and recognition ability is modulated by mean publication year, literary level and circulation frequency of the authors. Similarly, using a German test, Hug

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

et al. (2024) showed that differences in students' study programs, a choice that reflects personal interest, can influence the author recognition ability, with students of comparative literature scoring higher on the ART. This study also introduced genre as a potential variable, coding authors as fiction or nonfiction based on the genre associated with most of their work. However, Hug et al. (2024)'s study does not focus on the modulatory role of genre on print exposure, thus leaving its potential effects unexamined.

The most widely used English ART was developed nearly two decades ago for an American population (Acheson et al., 2008). Naturally, the literary landscape has changed since then. Because ART performance depends directly on the recognizability of the author names included, outdated item pools risk inaccurate print exposure scores. Moreover, the original validation was conducted primarily on student samples, limiting its suitability for capturing print exposure across the adult lifespan. As a result, the Acheson ART may no longer optimally capture individual differences in print exposure nowadays. The goal of this project is to develop an updated version of the test for American participants.

Older versions of the ART have exclusively used fiction authors. However, print exposure is not limited to literary fiction. Previous work has shown that time spent reading nonfiction is correlated with ART performance (Acheson et al., 2008). Furthermore, it can be argued that restricting the ART to only fiction puts non-fiction readers at a disadvantage and thus incorrectly represents their print exposure (Hug et al., 2024). We therefore include non-fiction authors in the updated version of the ART.

Another dimension that affects print exposure is literary level. Readers differ in their preferences for popular versus highbrow literature, and these preferences influence which authors they encounter (Grolig et al., 2020). We include both bestselling authors and recipients of prestigious literary awards, which allows the ART to sample different forms of cultural exposure, ranging from commercially popular authors to institutionally recognized authors. Balancing these dimensions can increase the sensitivity of the ART to variation in reading experience beyond canonical or school-mandated texts.

In sum, the current study presents and validates an updated version of the English Author Recognition Test for an American population. We construct a contemporary item pool based on recent bestsellers and award-winning authors, balance fiction and non-fiction genres, and distinguish between popular and highbrow literary levels. We evaluate this updated test psychometrically using item-response theory and compare it directly with Acheson et al. (2008)'s ART. In addition, we examine how participant-level variables such as age and education interact with item-level properties such as genre and literary level to modulate recognition probability. By updating the author pool and broadening the scope of the test, the goal is to provide a contemporary, valid and inclusive measure of print exposure.

Test creation

Candidate authors were first compiled from publicly available bestseller and award lists and then manually filtered by two researchers.[1] For the final version of the new ART, we selected 65 author names who have either published a bestselling book between the years 2012 and 2022 (according to the New York Times bestseller list, for which data was only available up until 2012), or have been awarded any prestigious international or American literary awards (i.e. the Booker Prize, the Nobel Prize for Literature and the National Book Award) for their work between the years 2002 and 2022. Since the New York Times bestseller lists are updated on a weekly basis and include multiple subcategories (such as fantasy, mystery, young adult etc.), the researchers selected bestselling authors with the highest number of appearances on all bestseller lists. Different time windows were used to balance recency and long-term cultural visibility: bestseller lists emphasize contemporary exposure, whereas award winners often remain recognizable over longer periods. Also, fewer authors are included in lists of literary award winners than in the weekly New York Times bestseller lists, meaning that we needed to use a bigger time window to collect enough author candidates. Additionally, only novelists were included on the author list (as opposed to poets and playrights), since we wanted our list to reflect experience with the literature that results in the largest amount of print exposure.

Following Hug et al. (2024), the author list includes writers from two genres: fiction and nonfiction. Genre was determined based on the dominant category of the author's published works and public reception, with authors assigned to the genre in which the majority of their well-known works appeared.

Following Grolig et al. (2020), the fiction authors were further subdivided according to their literary level: popular or highbrow. Popular authors were those who appeared on bestseller lists, while highbrow authors have received prestigious literary awards. This decision stems from the fact that literary prizes are awarded by literary gatekeepers to novels that display exceptional literary quality. Authors appearing both on bestseller and award lists were coded as highbrow. Disagreements about genre or literary level were resolved through discussion.

Table 1 summarizes the distribution of author names across conditions. We limited the nonfiction author names to a third of the test stimuli, since fiction authors have been shown to be a more sensitive measure for the ART (Wimmer & Ferguson, 2023; Acheson et al., 2008).

We also created a list of 65 foils, which were derived from

[1] Prior to the selection of the author list that was validated in the current contribution, a more comprehensive list of 99 authors was pilot tested on a sample of 100 native English speakers. Item-response theory and exploratory factor analyses performed on the test responses revealed potentially problematic items (in terms of difficulty and discriminative power) and informed the selection process of the final list. Full IRT and EFA results of the pilot test are available at `osf.io/axf3w`.

Table 1: Distribution of author names across conditions.

Condition	Authors
popular fiction	27 (41.5%)
highbrow fiction	18 (27.7%)
nonfiction	20 (30.8%)

census data of common first and last names for all the nationalities present in the test items (e.g. American, British, French, Japanese, Russian etc.). All generated names were checked against public databases to ensure they did not correspond to any real author or public figure. The foils were matched in terms of nationality and gender to the author names.

Test validation

Methodology

Participants. Two hundred native English speakers (age range: 18 – 50 years[2]; mean age = 34.11 years; SD = 8.84; 103 identified as male) registered as participants on the crowdsourcing website Prolific, took part in the study. All participants indicated that they were born in the United States and were currently living there. One hundred thirty one participants had completed post-secondary education (had earned an undergraduate degree, a graduate degree, or a doctorate) and 59 participants had completed high school, technical education or had no formal qualifications. Ten participants chose to not disclose their educational background.

Procedure. In order to assess the performance and suitability of the updated version of the ART, we added the author names from the Acheson et al. (2008) test to our validation. Five author names overlapped across the two tests, resulting in a total of 125 unique author items. We included additional foils such that the validation study consisted of an equal number of author names and foils. In other words, each participant completed an online ART consisting of the 65 author names we have selected, 65 author names from the Acheson et al. (2008) test and 125 additional non-author names that did not belong to any existing author.

The test was administered as an online survey hosted on Qualtrics (2026) and distributed via Prolific (2026). The test items were shown in alphabetical order to mirror the format of previous ART implementations. Names were presented one at a time as part of a multiple choice question that asked whether the name shown on the screen belonged to an author. The participants were instructed to answer by selecting "yes" or "no" and that they had to make their choice in maximum 6 seconds. The 6-second limit was chosen to encourage intuitive recognition rather than deliberate search strategies, following prior ART implementations (Scholman et al., 2024). After selecting their answer, they would move on to the next item. Participants were warned that they would receive a score penalty for each incorrect non-author selection. The administration of the updated ART and the Acheson et al. (2008) ART together took 8 minutes on average.

[2]Our decision to restrict the age range to 18–50 years was informed by the finding that print exposure effects stabilize after the age of 50 (Grolig et al., 2020).

Data preparation and analysis

The ART scores were calculated according to the standard method of subtracting the number of false alarms (foils incorrectly identified as authors) from the number of hits (correctly selected authors). Because the number of foils in this validation was doubled relative to a standard ART, false alarms were normalized by dividing by two to keep the penalty scale comparable to prior studies.

Following Scholman et al. (2024), we included two quality filters meant to identify participant data of very poor quality: a viable participant should not (i) time out (spend more than 6 seconds) on more than 33% of the trials and (ii) select more than 66% of the foil items. As a result of these two criteria, forty-eight participants were excluded from further analysis. This exclusion rate is comparable to other online ART studies.

The reliability of the new version of the ART was established through an item-response theory (IRT) analysis, which was performed using the `mirt` package (Rizopoulos, 2006) in R4.3.3 statistical software (R Core Team, 2024). To this end, we estimated a two-parameter logistic model (2-PL) which describes the items in terms of difficulty (*b-parameter*) and discriminative power (*a-parameter*) and determined its goodness of fit using AIC, BIC, and likelihood ratio tests.

The modulating effects of participant- and item-level variables on successful author recognition were examined through a generalized linear mixed-effects model (`lme4` package, Bates et al., 2015). The model included participant age, educational background, author literary level and genre as well as the interactions between participant variables and author variables as fixed effects. The model also included random intercepts for participants and items. The complete model structure is available in Table 3. The continuous variable age was centered. The categorical variables educational background, literary level and genre were effect coded, with lower education, popular author and fiction author as their respective reference levels (coded as -1).

Results

The updated ART vs. Acheson et al. (2008)'s ART. Table 2 shows the distribution of the hit rates, false alarms and ART test scores for both versions. Overall, participants recognized slightly fewer authors in the updated ART than in the Acheson ART (M = 26.17 vs. 28.93 hits), which resulted in a lower mean ART score for the new version (4.83 versus 7.59). This difference reflects the greater overall difficulty of the updated item set. The reliability of both tests was estimated using the split-half correlation and Cronbach's alpha. Both versions of the test had very good reliability scores (see Table 2). These results indicate that both test versions provide stable measurements of print exposure.

Table 2: Descriptive statistics for the two version of the ART. Rel.: Spearman-Brown corrected split-half reliability; α: Cronbach's alpha.

Test	Range	Mean	SD	Rel.	α
Hit rate: new	0 – 49	26.17	13.0		
Hit rate: Acheson	1 – 49	28.93	11.68		
False alarm (halved)	0 – 41.5	21.35	14.07		
new ART	-9 – 22.5	4.83	6.20	0.93	0.93
Acheson ART	-7.5 – 43.5	7.59	9.40	0.92	0.91

To examine the relationship between the newly developed ART and the established Acheson ART, participants' scores on both tests were compared using two complementary analytic approaches. First, a Pearson correlation was conducted to assess the degree of association between individual scores on the two measures. This analysis revealed a strong positive correlation between the new ART and the Acheson ART, r(150) = .69, $p < .001$, indicating that participants who scored higher on one test also tended to score higher on the other. Second, a paired-samples t-test was used to test whether the two ART versions differed systematically in overall difficulty. The analysis revealed a significant difference in mean scores (t(151) = 5.0,$p < .001$), with participants scoring significantly higher on the Acheson ART. Together, these results show that while the two tests strongly agree in ranking individuals, the updated ART is overall more challenging.

We do note, however, that the mean ART scores are relatively low compared to those reported in other studies (Acheson et al., 2008; Moore & Gordon, 2015; Scholman et al., 2020, 2024), which typically report average scores of around 15–22. To better understand this discrepancy, we compared our results with those of Scholman et al. (2024), who tested 239 American Prolific participants using the Acheson ART with foils from Moore & Gordon (2015). In their study, participants achieved a mean hit rate of 24.7 and a mean false-alarm rate of 8.1, with an average ART score of 19.7. In comparison, the hit rates in our study are similar across the updated ART, our Acheson implementation, and the implementation used by Scholman et al. (2024). Thus, the lower overall ART scores observed here are not driven by reduced recognition of real authors. Instead, the difference arises primarily from substantially higher false-alarm rates in our data.

We see two possible reasons for this. First, our validation study combined two ART versions into a single test, effectively doubling the amount of items. This increased length may have introduced greater fatigue, resulting in more liberal responding and therefore more false positives. However, if this were the case, we would expect to also see higher hit rates, which we do not find. A more plausible explanation is that our foils may have been more difficult to reject because they consisted of highly plausible, common names, which could have increased the likelihood of mistakenly endorsing non-authors. Crucially, this elevated false-alarm rate does not compromise the central goal of our study. Our primary interest lies in the recognition of real author names and validation of those items. The hit rates in this validation study are comparable to those in previous implementations, which provides insight into the validity of the author names in the test. We will return to the issue of foil difficulty in the Discussion.

Reliability of the updated ART. The two-parameter logistic (2-PL) model estimated on the 65 items of the new version of the ART demonstrated excellent fit to the data: M2(1952) = 2618.72, $p < .001$, RMSEA = .042 (90% CI [.038, .046]).[3] Both simpler (1PL) and more complex (3PL) models were also estimated and compared to the 2PL model. Likelihood ratio tests indicated that the 2-PL model provided significantly better fit than the 1-PL model, $\chi^2(64) = 294.68, p < .001$, while the 3-PL model did not significantly improve upon the 2-PL model, $\chi^2(65) = 4.45, p = 1.00$. Thus, the 2-PL model was retained as the most parsimonious and well-fitting representation of the data. A second 2-PL model was similarly estimated and fit for the Acheson et al. (2008) items.

A 2-PL model describes a test across two dimensions: item discrimination (a-parameter) and item difficulty (b-parameter) in relation to the underlying ability measured by the test (i.e., print exposure in our case). Figure 1 illustrates the distribution of these for the two test versions. The item difficulty parameter (b) indicates the level of ability required for an individual to have a 50% probability of correctly recognizing an author. Lower (including negative) b-values correspond to easier items, whereas higher b-values correspond to more difficult items. In our case, the two tests show differences in b-parameter range: the new version of the ART has a smaller range (-2.99 – 2.01) compared to the Acheson version (-5.56 – 5.97): Levene's test revealed that the variance of the b-parameters of the two tests was significantly different (F(1, 127) = 4.56, p = .035), with the Acheson difficulty parameters displaying greater variability (Acheson: SD = 1.59, new: SD = 0.79). However, a Welch's t-test indicated that the mean item difficulty of the two tests was not significantly different (new: 0.37, Acheson: 0.15; t(94.23)=0.99, p = 0.325). These results show that the items in the new version are more homogeneous in terms of difficulty, whereas the Acheson ART spans a broader spectrum from very easy to very difficult items.

The discrimination parameter (a) represents the slope of an item's logistic function at the point of inflection, whose location is determined by the b-parameter. It captures how strongly an item differentiates between participants with lower versus higher levels of print exposure. Higher a-values therefore indicate items that contribute more strongly to separating individuals along the latent ability dimension. In our case, the results indicate that the distributions are compara-

[3]The significant M2 statistic usually indicates that the model does not fit the data perfectly, which is prevalent in big sample sizes (Maydeu-Olivares & Joe, 2006). In contrast, a root-mean-square error of approximation (RMSEA) value of $< .05$ signals that the model is a good fit for the data (Hu & Bentler, 1999). Taken together, the results show that the model is a good fit, although not perfect.

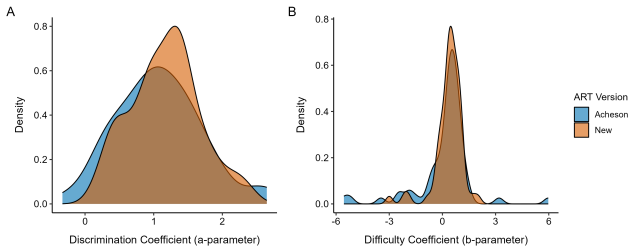

Figure 1: Density distributions of the a- and b- parameters of the 2-PL models estimated for the Acheson ART and the new ART

ble between the two tests (Kolmogorov-Smirnov test: D = 0.13, p = 0.53). The range of the new version's a-parameter is slightly shifted to the right (new: 0.17 – 2.39; Acheson: -0.33 – 2.64). However, there was no significant difference between the two tests in terms of the variance (F(1, 127) = 1.72, p = 0.19) or the means (t(127) = 0.89, p = 0.37).

The effect of item- and participant-level variables

Table 3 presents the GLMER results for testing the effect of item- and participant-level variables on the accuracy of author recognition. Main effects revealed that more educated participants showed higher recognition accuracy overall. Popular authors were recognized more accurately than highbrow authors, and fiction authors were recognized more accurately than nonfiction authors. Participant age did not significantly modulate print exposure. These results indicate that both reader characteristics and author characteristics systematically modulate recognition performance.

Two significant interaction effects emerged between educational level and both author variables. Specifically, the advantage for popular over highbrow authors was reduced among more educated participants, indicating that higher-educated readers show relatively better knowledge of highbrow authors, thereby narrowing the gap between popular and highbrow recognition. Similarly, higher-educated readers also showed better knowledge of nonfiction authors compared to lower-educated individuals, thus reducing the observed genre effect.

Random effects revealed substantial individual differences in overall author knowledge and moderate variability in author recognizability. At the participant level, individuals with higher overall recognition ability showed stronger advantages for popular authors (r = .45) and fiction authors (r = .69), suggesting that variation in print exposure is particularly driven by familiarity with mainstream fiction. At the author level, more recognizable authors showed smaller education effects (r = -.52), indicating that highly famous authors tend to be known across educational groups, whereas less recognizable authors are disproportionately known by more educated participants.

Table 3: Regression coefficients and test statistics from the generalized linear mixed-effects model. Significant predictors ($p < 0.05$) are marked with *. Abreviations: **Edu** = educational background, **Lit** = literary level, **Part** = participant, **Int** = Intercept

	β	*SE*	*z*	*p*
Intercept	-.78	0.13	-5.85	< .001*
MAIN EFFECTS				
Age	-.02	0.01	-1.24	.22
Edu	.29	.11	2.60	.009*
Lit	-.26	.07	-3.46	< .001*
Genre	-.18	.08	-2.34	.02 *
INTERACTIONS				
Age:Lit	-.001	.004	-.36	.71
Age:Genre	-.002	.004	-.57	.57
Edu:Lit	.11	.04	2.73	.006*
Edu:Genre	.08	.04	2.03	.04*
RANDOM EFFECTS				
	Variance	SD	Corr	
Participant				
Intercept	1.42	1.19		
Literary level	.05	.21	.45 (Int)	
Genre	.05	.21	.69 (Int), .60 (Lit)	
Authors				
Intercept	.25	.50		
Edu	.02	.14	-.52 (Int)	
GLMER model structure				
Fixed effects: $Age + Edu + Lit + Genre + Age:Lit + Age:Genre + Edu:Lit + Edu:Genre$				
Random effects: $(1 + Edu\|Author) + (1 + Lit + Genre\|Part)$				

Discussion

The present study investigated print exposure differences in a crowdsourced sample of 18- to 50-year old readers with various educational backgrounds. Our study builds upon the well-known Author Recognition Task developed by Acheson et al. (2008) for American English speakers, using the insights of Grolig et al. (2020) about how item- and participant-level variables influence an individual's ability to correctly identify author names.

The result of the current study is the creation of an updated, balanced, and inclusive version of the ART that can be administered to English L1 adult speakers in the United States of more various ages and educational backgrounds than the classic student population. Our item set generated hit rates similar to those of different implementations of the Acheson item set. Moreover, our item-response analysis revealed that the two author lists were comparable in terms of difficulty and discriminative power, with the only difference being that the Acheson author list includes items with more extreme dif-

ficulty values. In particular, the Acheson author list contains several items with very low difficulty scores (between -5 and -2), such as Ernest Hemingway, F. Scott Fitzgerald and George Orwell. These authors are both known for their writing, as well as for the adaptation of their work into films and television series or for their influence on social and political discourse (in the case of Orwell). Therefore, it is possible that individuals with very low print exposure are able to recognize their names. Furthermore, the low item discrimination scores of these three example authors (Hemingway: 0.21; Fitzgerald: 0.19; Orwell: 0.38) indicate that their inclusion is not very informative of different levels of print exposure. In our implementation of the two ART tests, the slight score advantage shown by the Acheson author list is therefore driven by a few low difficulty and uninformative items.

We believe that our author list is a suitable alternative to Acheson's, which has been in use for almost 20 years and has therefore become outdated. Additionally, the Acheson ART is a widely used measure for individual differences in print exposure for English speakers, which means that with every new study using its item list to estimate print exposure, the possibility that the participant sample has received prior exposure to these test items increases. We offer a new set of author names that has been constructed to fit more heterogeneous participant samples and includes prestigious and commercially successful authors as well as fiction and non fiction writers to create a measure that is inclusive of different reading patterns.

One notable disadvantage of our implementation is the creation of foil items. There is little precedent for foil creation strategies, since most pre-existing literature on ART development focuses its reporting on the process of author name selection, while dedicating little space to foils. Most notably, Brysbaert et al. (2020) selected their foils from pre-existing name lists of people that were unlikely to be known, such as participants in nonprofessional running competitions or World War I victims. We decided on an alternative approach, in order to broaden the range of available foil creation strategies. The intuition behind this choice was that by creating new combinations of common first and last names, we will derive names that are less distinctive and less likely to attract attention. However, in practice, we did not account for the possibility that the familiarity of these names would make participants more likely to select them. Among the foils that resulted in the highest false alarm rates, a common characteristic seems to be that the last names are reminiscent of public figures: e.g. Albert Dean (false alarm rate 44%) shares the same last name as the actor James Dean, and Judy Hawkins (false alarm rate 48%) shares the same last name as the physicist Stephen Hawkins. Another common trait of the most frequently chosen foils is the fact that they are English sounding (e.g. Richard Blair, Ryan Hart, Madeline Burke). In comparison, the non-English sounding foils had low false alarm rates ($< 30\%$). The high selection rate of the English sounding foils is likely due to both the American cultural hegemony as well as participants living in the United States and thus having knowledge of more American sounding famous names. This extensive exposure can potentially lead to the creation of false connections. We therefore recommend future implementations of the ART use a different set of foils.

The secondary aim of this study was to further explore the influence of both participant-level variables (age and educational background) as well as item-level variables (literary level and genre) on print exposure. Only educational background was a significant participant-related predictor for author recognition accuracy. Higher educated individuals had better overall accuracy, and were also specifically better at identifying highbrow and nonfiction authors than individuals with a lower educational background.

At the same time, recognition ability did not vary significantly between different participant ages, regardless of the genre and literary level of the author. This result contrasts in particular with the findings of the German ART study (Grolig et al., 2020), which observed significant interaction effects of age and literary level, but no main effect of age. These conflicting results could be explained through sample differences: in the German ART study, the two thirds of the participants were recruited from book fairs, which are places frequented by people with an active interest in reading. Therefore, the sample for the German ART is biased towards better ART performance. Meanwhile, our crowdsourced sample was more random, since we did not sample from audiences that are specifically likely to have a personal interest in reading. Additionally, Grolig et al. (2020) showed that older participants were more likely to recognize classic authors (mean publication year 1965) than recent authors (mean publication year 2015). In comparison, our ART did not include older authors, due to limitations of archival access to the New York Times bestseller lists. Therefore, it is plausible that the inability to replicate the age effect was due to both author recency of our item list as well as differences in participant sample characteristics.

Genre and literary level were significant predictors of author recognition accuracy, with popular and fiction authors being recognized more accurately. These results indicate that accounting for such item traits as literary level and genre during author list construction introduces items that manage to discriminate between different levels of print exposure. More clearly, highbrow and nonfiction authors are items that separate individuals with a higher level of ability, while popular and fiction authors have a higher likelihood to be correctly recognized by individuals with less print exposure.

To conclude, our results show both that participant sample heterogeneity as well as variation in author characteristics can influence author recognition performance, and that all these factors should be taken into account when designing or choosing the author list for a print exposure measure.

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*(1), 278–289.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Brysbaert, M., Sui, L., Dirix, N., & Hintz, F. (2020). Dutch author recognition test. *Journal of Cognition*, *3*(1).

Chen, S.-Y., & Fang, S.-P. (2015). Developing a chinese version of an author recognition test for college students in taiwan. *Journal of Research in Reading*, *38*(4), 344–360.

Gedik, T. A. (2024). Development of the turkish author recognition task (tart) and the turkish vocabulary size test (turvost). *SN Social Sciences*, *4*(8), 151.

Grolig, L., Tiffin-Richards, S. P., & Sascha, S. (2020). Print exposure across the reading life span. *Read Writ*, *33*, 1423–1441. Retrieved from https://doi.org/10.1007/s11145-019-10014-3

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55.

Hug, M., Jarosch, J., Eichenauer, C., Penella, S., Kretzschmar, F., & Nicklas, P. (2024). Some students are more equal: Performance in author recognition test and title recognition test modulated by print exposure and academic background. *Behavior Research Methods*, *56*, 6004–6019. Retrieved from https://doi.org/10.3758/s13428-023-02330-y

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*(3), 273–284. Retrieved from https://doi.org/10.1111/j.1467-9817.2008.00371.x

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. doi: 10.1007/s11336-005-1295-9

McCarron, S. P., & Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native english speakers? an item response theory analysis. *Behavior Research Methods*, *53*(5), 2226–2237.

Mol, S., & Bus, A. (2011, 01). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*, 267-296. doi: 10.1037/a0021890

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109.

Prolific. (2026). *Prolific: A platform for online research.* Retrieved from https://www.prolific.com/ (Accessed: 2025-01-31)

Qualtrics. (2026). *Qualtrics survey software.* Retrieved from https://www.qualtrics.com/ (Accessed: 2025-01-31)

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1-25. Retrieved from https://www.jstatsoft.org/v17/i05/ doi: 10.18637/jss.v017.i05

Scholman, M., Demberg, V., & Sanders, T. J. M. (2020). Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, *57*(10), 844–861. Retrieved from https://doi.org/10.1080/0163853X.2020.1813492 doi: 10.1080/0163853X.2020.1813492

Scholman, M., Marchal, M., & Demberg, V. (2024). Connective comprehension in adults: The influence of lexical transparency, frequency, and individual differences. *Discourse Processes*, *61*(8), 381–403.

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

Stoops, A., & Montag, J. L. (2023). Effects of individual differences in text exposure on sentence comprehension. *Sci Rep*, *13*(16812). doi: https://doi.org/10.1038/s41598-023-43801-8

Wimmer, L., & Ferguson, H. J. (2023). Testing the validity of a self-report scale, author recognition test, and book counting as measures of lifetime exposure to print fiction. *Behavior Research Methods*, *55*, 103–134. Retrieved from https://doi.org/10.3758/s13428-021-01784-2

Zufferey, S., Degand, L., Popescu-Belis, A., & Sanders, T. J. M. (2012). Empirical validations of multilingual annotation schemes for discourse relations. In *Proceedings of the 8th joint iso-acl/sigsem workshop on interoperable semantic annotation (isa-8)* (p. 77-84). Jeju Island, Korea.

Zufferey, S., & Gygax, P. (2020). "roger broke his tooth. however, he went to the dentist": Why some readers struggle to evaluate wrong (and right) uses of connectives. *Discourse Processes*, *57*(2), 184–200. Retrieved from https://doi.org/10.1080/0163853X.2019.1607446 doi: 10.1080/0163853X.2019.1607446