# Estimating Type of Print Exposure across Aging through Author Production


**Mengyang Qiu (mengyangqiu@trentu.ca)**
Department of Psychology, Trent University
Peterborough, ON K9L 0G2, Canada

**Nichol Castro (nicholca@buffalo.edu)**
Department of Communicative Disorders and Sciences, University at Buffalo
Buffalo, NY 14214, United States

**Brendan T. Johns (brendan.johns@mcgill.ca)**
Department of Psychology, McGill University
Montreal, QC H3A 1G1, Canada





## Abstract

This study introduces a novel approach for quantifying individual differences in print exposure through the integration of distributional semantics with the Author Production Test (APT). By employing the Universal Sentence Encoder to generate vector representations of authors from their works, we constructed 'participant vectors' reflecting the aggregated author vectors individuals produced in the APT and 'genre vectors' capturing the representative characteristics of each literary genre. By analyzing the cosine similarities between participant and genre vectors, we objectively estimated individuals' genre preferences. The results demonstrated a significant correlation between these objective measures and self-reported genre preferences, particularly for older frequent readers, highlighting the method's effectiveness. Our findings offer a promising avenue for the objective measurement of print exposure, with potential implications for developing personalized models of lexical behavior.


**Keywords:** cognitive aging; print exposure; distributional semantics

## Introduction

Studies in healthy aging often find that older adults exhibit decreased performance in various cognitive tasks compared to their younger counterparts. While many theories have posited that this underperformance is due to a decline in older adults' cognitive systems (e.g., Naveh-Benjamin, 2000; Hasher & Zacks, 1988), recent research points to the role of accumulated knowledge in accounting for these age-related behavioral variations (Ramscar et al., 2014; Qiu & Johns, 2020).

Aging is not merely associated with the accumulation of knowledge, but also with the types of knowledge (e.g., everyday terms versus occupation specific terms) acquired throughout an individual's life. Differences in language exposure arise from various socioeconomic, educational, and generational factors (Wulff et al., 2019). Even within the same age group, individuals may exhibit diverse linguistic knowledge, such as varying levels of sensitivity with word meaning and usage, resulting from their unique language exposure.

A recent analysis of variance in written language by Johns and Jamieson (2018) provides evidence supporting the significant role of language exposure in cognitive research. The study analyzed a large collection of fiction books, spanning seven genres and hundreds of authors, to discern patterns and disparities in word usage among authors, within and across genres, and throughout different time periods. As expected, books across different genres (e.g., science fiction vs. romance) were less similar than those in the same genre, and books from the same time period were more similar than books from different periods. Notably, the study revealed considerable individual variability in language use among authors, with the highest similarity observed in the within-author comparisons. Since people's reading and media consumption habits tend to be selective rather than random, accounting for personalized language experiences becomes crucial in understanding individual differences, including age-related variances, in language processing and associated cognitive tasks.

Further research by Johns and colleagues, primarily focusing on the group level (e.g., American English speakers vs. British English speakers), has reinforced the importance of considering language experiences when explaining behavioral data (e.g., Johns & Jamieson, 2019; Johns et al., 2019; Taler et al., 2020). However, an essential question remains: how can we measure these differences at the individual level? Modeling language exposure at the individual level not only allows for a more precise understanding of psycholinguistic phenomena, but also has important implications in fields like education and neuropsychology. By tailoring teaching materials and assessing items to an individual's specific language experience, we could potentially optimize language acquisition and neuropsychological evaluations.

Language exposure encompasses diverse sources, including books, movies and TV shows, social media, everyday spoken communication, etc. As a first attempt to model language exposure, our focus was on reading, also termed as print exposure. Despite the rise in social media usage negatively affecting reading volume (Twenge et al., 2019), reading remains an important source of knowledge accumulation, as books contain more diverse vocabulary (Johns et al., 2020) and complex grammatical structures (Roland et al., 2007).

Traditionally, print exposure was measured using self-reported reading questionnaires (McGeown et al., 2015) or reading diaries (Anderson et al., 1988). However, these methods can be subjective (e.g., over-estimating reading preference given the socially desirable nature of reading) and labor-

Accepted for presentation as a remote talk at CogSci 2024

intensive. A more objective measure of print exposure is the Author Recognition Test (ART), where participants are required to discriminate between names of real author and foils presented in a list (Stanovich & West, 1989; Acheson et al., 2008). The ART has been established as a reliable and valid measure for print exposure, particularly from childhood through early adulthood (Mol & Bus, 2011).

In order to estimate the type of print exposure, one possible approach involves selecting authors from various genres. By analyzing how accurately participants recognize authors from different genres, we can infer their exposure to and familiarity with these genres. However, recent studies have shown that selecting author items with appropriate difficulty to capture the range of print exposure and to drive individual differences in a given population is particularly challenging (Brysbaert et al., 2020). Furthermore, the ART, though validated in certain populations, may show reduced effectiveness when applied to different populations (McCarron & Kuperman, 2021) or during different time periods (Moore & Gordon, 2015), thus posing challenges for its use in age-comparative research.

Therefore, in the current study, we proposed an alternative approach: the Author Production Test (APT), which can be viewed as a semantic fluency task on the category of fiction book authors. By focusing on author name production, the APT can potentially bypass the challenges of item selection in the ART. To estimate the type of print exposure, we can categorize the genres of the authors that each individual produce. For instance, if a participant produces ten science fiction authors and two romance authors, it might indicate that their reading experiences predominantly lie within the science fiction genre. This, in turn, could be used to understand their unique print exposure and to construct individualized corpora to model their behavioral data.

In order to more accurately capture genre information, we utilized the Universal Sentence Encoder (USE; Cer et al., 2018)[1], to derive vector representations of authors based on their concatenated book descriptions[2], obtained from the UCSD Goodreads dataset (`https://mengtingwan.github.io/data/goodreads.html`; Wan et al., 2019). This distributional semantics approach, demonstrated to be effective in recent semantic fluency studies (e.g., Taler et al., 2020), overcomes the limitations of of manually coding authors into one genre or another, which may not represent the full semantic and thematic breadth of an author's work. By accumulating the vectors corresponding to the authors listed by an individual (hereafter referred to as the 'participant vector'), this approach allows for the detection of subtle differences in genre preference across individuals. By comparing each participant vector to the most 'typical' or 'representative' vector for each genre (hereafter the 'genre vector'), we can objectively estimate individuals' genre preferences, as detailed in the next section.

The ultimate goal of this study is to validate a promising new approach for measuring the type of print exposure. By employing distributional semantics and leveraging the APT, we propose a quantitative representation of individuals' reading preferences. If our approach proves successful, it will offer an easy yet effective tool for investigating the subtle and complex landscape of individual reading habits and preferences, ultimately paving the path for a more granular modeling of lexical behavior.

## A Distributional Estimation of Genre Preferences

Given that authors often write in multiple genres and a single work can carry multiple genre tags on platforms like Goodreads and Amazon, a critical question emerges: How do we estimate the genre vectors? Johns et al. (2020) addressed this by categorizing authors according to their dominant genre, determined by the primary genre tag assigned on both Goodreads and Amazon. While this methodology may not capture the full genre spectrum of each author, it is worth noting that most genres in their study included hundreds of authors and thousands of books. Therefore, by comparing all the author vectors within a particular genre, it is likely to uncover recurrent patterns and commonalities that reflect the essence of that genre. Through this collective analysis, we can infer a representative vector of that genre.

To identify each genre vector, we utilized the concept of "medoid" rather than simply summing the author vectors within each genre. A medoid, or medoid vector, is a well-established statistical measure representing the most centrally located point within a dataset (Struyf et al., 1997). Thus, it effectively captures the 'typical' features of authors within a specific genre, offering a point of reference for understanding the genre's characteristics.

The process of finding the medoid for a genre involved the steps below. First, we segregated the authors according to their respective genres, obtained from Johns et al. (2020). Second, within each genre, the standard pairwise Euclidean distances were calculated between all the author vectors, resulting in a distance matrix[3]. Finally, by summing the distances from each author vector to all others, we determined the vector with the smallest total distance. This vector, known as the medoid, represents the most centrally located point

[1]The USE is designed to encode sentences or paragraphs into fixed-size vectors (512 dimensions), rather than focusing on single words like many other distributional semantics models. This capability makes it especially suitable for the current study, where book descriptions (paragraphs) were used to capture the semantic/thematic information of authors.

[2]Ideally, we would obtain the complete collection of works for all the authors produced. However, given the vast number of authors and the sheer volume of literary works, obtaining such a comprehensive dataset is impractical. Our exploratory results showed that the vector representations generated from book descriptions using the USE model exhibit a meaningful level of correspondence with those derived from complete books using the bag-of-words model. In other words, book descriptions appear to provide a reasonably accurate, albeit condensed, representation of the linguistic patterns and thematic content present in an author's full literary works.

[3]Given that the vectors derived from the USE are approximately normalized, calculating the Euclidean distances approximates cosine similarity, as reflected in the relationship $d = \sqrt{2(1-\cos(\theta))}$. The use of Euclidean distance was favored as it is the default metric implemented in the Python `scikit-learn` library.

within the high-dimensional space of that genre, and serves as a robust summary of the genre's overall characteristics.

While we aimed to find the representative vector for each genre, our objective for the participant vector was to capture the holistic representation. Thus, for each participant vector, we summed all the corresponding author vectors produced in the APT. In addition, due to different participants producing varying numbers of authors, we performed unit normalization on the summed vectors, which scales the vectors to have a consistent magnitude while preserving their directions (i.e., distributional information).

To objectively estimate genre preferences or the type of print exposure, we employed the commonly used cosine similarity between each participant vector and various genre vectors (hereafter referred to as 'objective genre preferences'). The cosine similarity yields a value between $-1$ and 1, where values close to 1 indicate a strong similarity or preference for a specific genre, values close to $-1$ indicate dissimilarity or aversion to a genre, and values around 0 suggest neutrality or no strong inclination towards a particular genre.

## Method

### Participants

A total of 324 younger (18–30 years old) and 318 older (50–70 years old) adults were recruited from Prolific.co (Palan & Schitter, 2018). Participants were pre-screened by Prolific to be native speakers of American English, have normal or corrected-to-normal vision, and be right-handed. Each participant was paid \$2 (at an average rate of \$10.00/hr) for completing the study. The whole study, including a language background survey and the APT, took less than 15 minutes to complete.

### Materials and Procedure

The language background survey was designed to collect information about participants' personal backgrounds and language experiences, and it was also intended for use in future studies. Among all the questions included, two specific survey questions were directly related to the current study. One of the questions was a multiple choice question asking, 'How many fiction books do you read in a year?' The available choices were: None, Fewer than 5, Between 5–10, Between 10–20, and More than 20. The other question was a Likert scale question designed to evaluate participants' self-reported preferences for popular genres of fiction books. On this scale, 1 represented the least preference and 7 indicated the most preference. The genres assessed included crime/mystery, fantasy, horror, literature, romance, science fiction, and thriller.

In the APT, participants were asked to list (type in) authors of fiction books they could recall reading. Participants were instructed to spend approximately 3 minutes on the task, but they were permitted to proceed after 1.5 minutes. To minimize the influence of cognitive control during author name retrieval, particularly among older adults (e.g., the added pressure of having to remember previously mentioned names; Mata & von Helversen, 2015), all the typed-in author names would remain visible on the screen.

## Results

Data from 24 older and 15 younger participants were excluded because they submitted the survey without completing the author production task, leaving data from 294 older and 309 younger participants for further analyses. Data processing, visualization and analysis were performed using R (v4.2.2; R Core Team, 2022) and Python (v3.10.9).

Author names produced were checked for spelling and normalized by computing the Levenshtein edit distance (`https://rapidfuzz.github.io/Levenshtein/`) with author names in the UCSD Goodreads dataset (e.g., *JRR Tolkien*, *Tolkien*, and *J. R. R. Tolkein* corrected to *J.R.R. Tolkien*). Names with greater edit distance (with an edit score below 90, where 100 indicates a perfect match with a name in the dataset) were manually reviewed. Names that could not be found in the dataset or on Goodreads.com (less than 10) were removed from analyses.

### Descriptive Statistics

Figure 1 presents the distribution of the average number of fiction books read per year among younger and older adults. The figure reveals significant variation both within and between the age groups. To facilitate a more straightforward analysis, we categorized participants based on their fiction reading habits. Participants who reported reading no fiction books or fewer than five fiction books per year were classified as infrequent readers, while those who reported reading 5 to 10, 10 to 20, or more than 20 books per year were classified as frequent readers. This resulted in a $2 \times 2$ design, with age group (younger adults vs. older adults) and fiction reading frequency (frequent readers vs. infrequent readers) as the factors. Table 1 presents a summary of the demographic statistics for each group of participants.

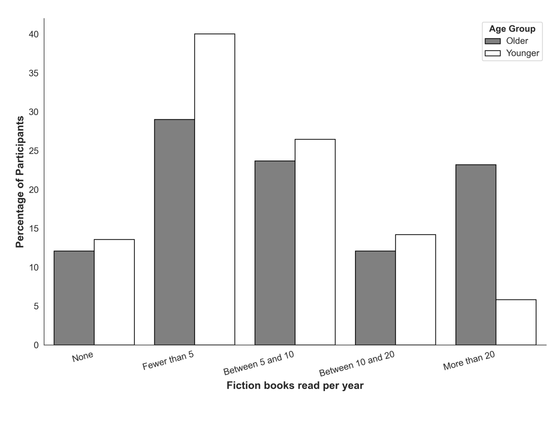


Figure 1: Distribution of the average number of fiction books read per year among younger and older adults.

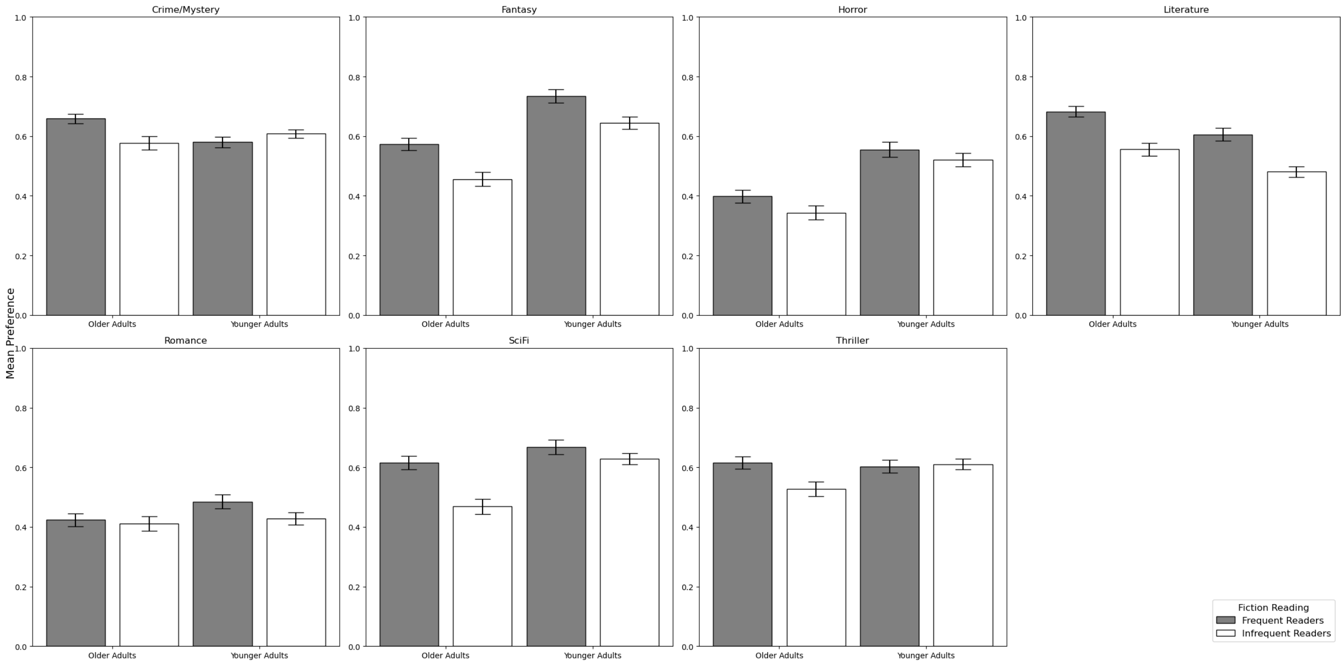


Figure 2: Mean Preference for Different Genres by Age Group and Fiction Reading Frequency. The preferences were scaled from 0 to 1 from the original Likert scale of 1 to 7.

In terms of genre preference, variations were observed across age groups and reading frequencies, as visualized in Figure 2. The results largely align with intuitive expectations about genre preferences among different demographic groups (e.g., older frequent readers displayed the highest preference for the literature genre, while younger frequent readers had a particular liking for the fantasy genre).

Table 1: Demographic statistics (reported as mean ± standard deviation).

| Age Group | Fiction Reading | $N$ | Age | Education (in years) |
|---|---|---|---|---|
| Older Adults | Frequent readers | 169 (120 female) | $57.28 \pm 5.84$ | $15.75 \pm 2.13$ |
| | Infrequent readers | 125 (87 female) | $57.624 \pm 5.33$ | $15.22 \pm 2.12$ |
| Younger Adults | Frequent readers | 137 (89 female) | $24.77 \pm 3.73$ | $14.27 \pm 2.28$ |
| | Infrequent readers | 172 (90 female) | $23.82 \pm 3.72$ | $14.24 \pm 2.09$ |

## Visualization and Analysis of Genre Vectors

The authors within each genre were identified by finding the intersection of authors listed in Johns et al. (2020) and the UCSD Goodreads dataset using a simple string matching technique. Johns et al. (2020) provided the genre tagging information for each author, while the UCSD Goodreads dataset furnished the book description data.

To visualize the spatial arrangement of authors within each genre, we utilized t-SNE (t-distributed Stochastic Neighbor Embedding), a machine learning algorithm that is particularly well-suited for visualizing high-dimensional data (Van der Maaten & Hinton, 2008). The resulting t-SNE plot (Figure 3) revealed that authors within each genre tended to locate closer to each other, a finding that aligns with the results of Johns and Jamieson (2018), and offers an intuitive confirmation of the genre tagging provided by Johns et al. (2020).

Since each medoid corresponds to an actual author vector, it offers the additional advantage of being able to identify the specific author who best represents the characteristics of each genre. The typical author for each genre, along with their demographic information and the total number of authors used to estimate the medoid for each genre, is provided in Table 2.

Table 2: Number of Authors and Representative Author Characteristics for Each Genre.

| | | Representative Author (Medoid) | | | | |
|---|---|---|---|---|---|---|
| Genre | No. of Authors | Name | Gender | Date of Birth | Average Rating | No. of Ratings |
| Crime/Mystery | 338 | Robert B. Parker | Male | 1932 | 3.92 | 314998 |
| Fantasy | 222 | A. A. Attanasio | Male | 1951 | 3.87 | 6787 |
| Horror | 72 | Simon Clark | Male | 1958 | 3.84 | 17371 |
| Literature | 339 | Sebastian Faulks | Male | 1953 | 3.83 | 108203 |
| Romance | 536 | Jo Goodman | Female | 1953 | 3.81 | 13159 |
| SciFi | 334 | Greg Bear | Male | 1951 | 3.84 | 452665 |
| Thriller | 154 | Jack Higgins | Male | 1929 | 3.92 | 138754 |

## Correlational and Regression Analyses

A preliminary correlational analysis was conducted to explore the relationship between the objective genre preferences and self-reported/subjective genre preferences, moderated by age and reading frequency. As shown in Figure 4, a general pattern emerged across all genres. That is, older frequent readers exhibited the highest correlation between their objective and subjective genre preferences. These findings suggest that our approach effectively captures the type of print exposure, particularly for older frequent readers who have accu-

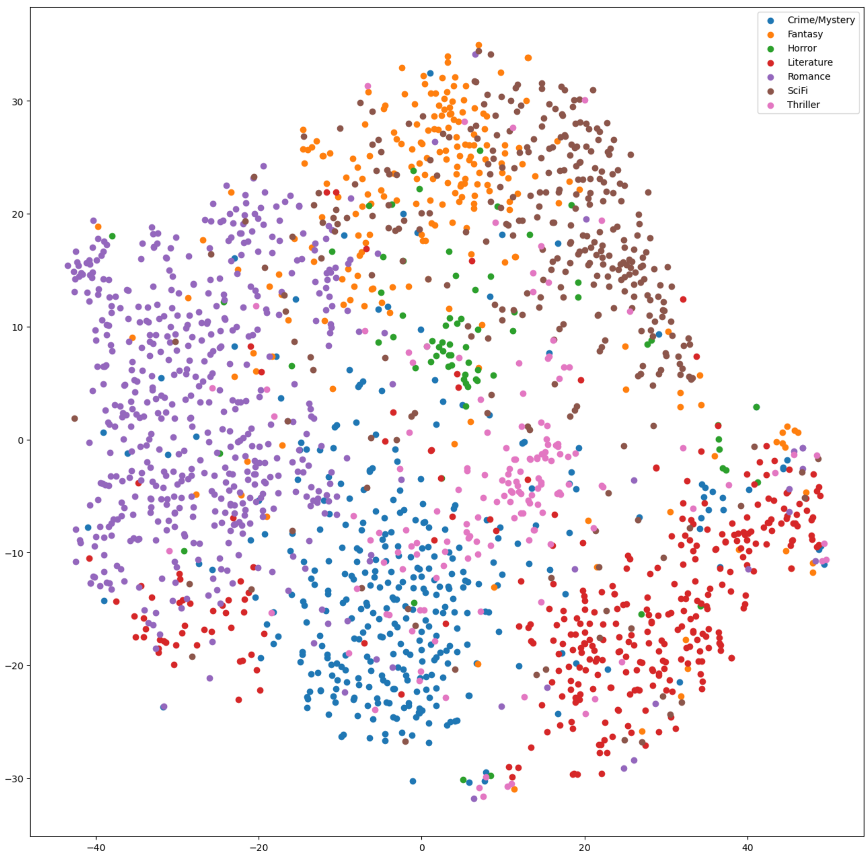


Figure 3: t-SNE plot representing the spatial arrangement of authors within each genre in a 2-dimensional space. Each dot in the plot corresponds to an author, and different colors represent different genres.

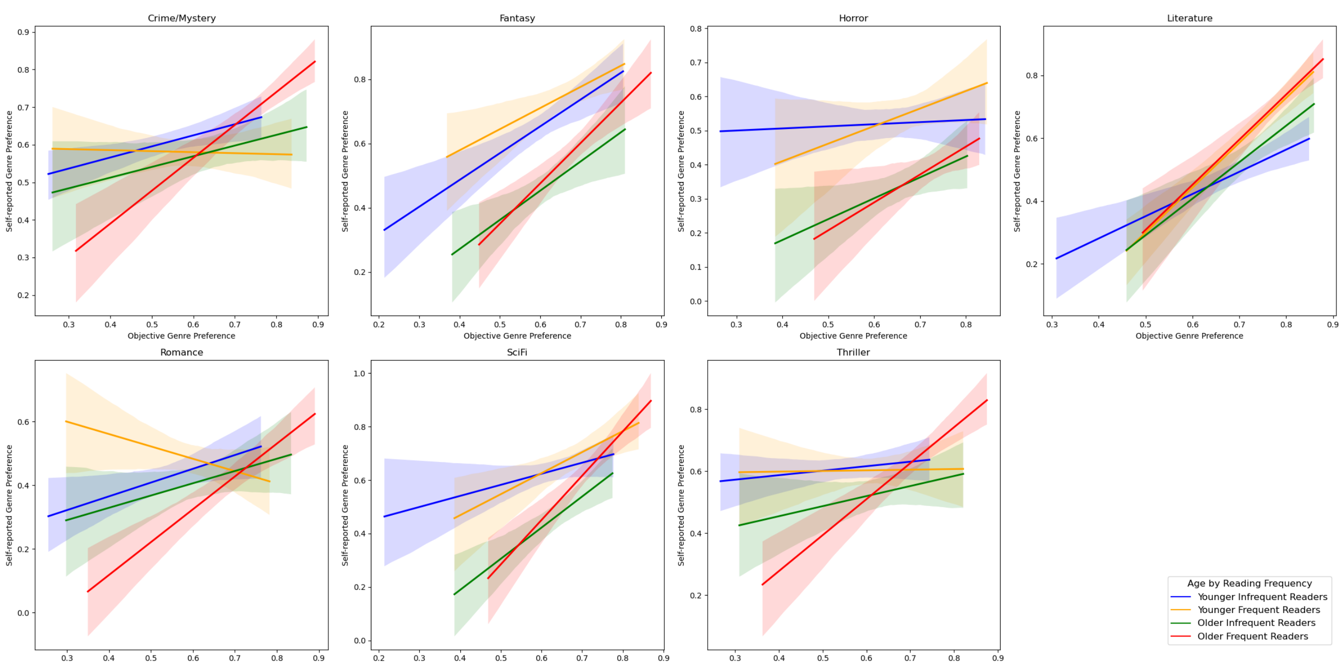


Figure 4: Correlation between objective genre preferences and self-reported genre preferences, moderated by age and reading frequency groups.

mulated a larger amount of print exposure.

To rigorously investigate the interaction of age group and reading frequency with the objective genre preferences in predicting subjective genre preferences, we then utilized a mixed-effects regression model, where participants and genres were treated as random intercepts. The results revealed significant main effects of age group, reading frequency, and objective genre preferences. Specifically, younger readers reported a higher genre preference ($\beta = 0.63, SE = 0.10, t = 6.41, p < .001$). Infrequent readers also showed a higher self-reported genre preference ($\beta = 0.46, SE = 0.10, t = 4.70, p < .001$). Additionally, a higher objective genre preference was associated with a higher self-reported genre preference ($\beta = 1.71, SE = 0.10, t = 17.00, p < .001$).

Two interaction effects showed significant relationships. First, the interaction between objective genre preferences and age group was significant, suggesting that the positive relationship between objective and self-reported genre preferences was less pronounced for younger participants compared to older ones ($\beta = -0.73, SE = 0.14, t = -5.13, p < .001$). Second, the interaction between objective genre preferences and reading frequency was also significant. Specifically, the relationship between objective and self-reported genre preferences was less strong for infrequent readers compared to frequent readers ($\beta = -0.67, SE = 0.14, t = -4.74, p < .001$). However, the two-way interaction between age group and reading frequency was not significant ($\beta = -0.17, SE = 0.13, t = -1.36, p = .17$). Similarly, the three-way interaction among objective genre preferences, age group, and reading frequency did not yield a significant result ($\beta = 0.20, SE = 0.19, t = 1.02, p = .31$).

In conclusion, the regression results further corroborate the observations drawn from the preliminary correlational analysis. The data illustrates a clear pattern: both age and self-reported reading frequency have significant interactions with objective genre preferences in predicting self-reported genre preferences. Older readers and those who read more frequently exhibited a stronger association between their objective and subjective measures of print exposure.

## Discussion

This study, joining a growing body of research underscoring the importance of individual differences in language exposure (Johns & Jamieson, 2018; Johns et al., 2019, 2020), proposes an innovative approach to capturing these differences at a granular level. We sought to quantify the type of print exposure, which has traditionally been a challenging aspect to measure, through a distributional analysis of the APT. Our results substantiate the effectiveness of our methodology, highlighting the validity of each stage of our approach: ranging from the derivation of genre and participant vectors to the inference of genre preferences through the cosine similarity of these vectors. The significant correlation between the cosine similarity measure and self-reported genre preference, especially among older frequent readers, indicates that our approach could objectively quantify an individual's genre preference and the extent of their exposure to different types of print. Moreover, it provides empirical support for our proposition that there exists an interaction between the amount and type of print exposure, with a more extensive reading habit leading to a more accurate self-understanding of one's genre preference.

Interestingly, we found that both infrequent readers and younger adults reported higher genre preference ratings. These findings resonate with previous research indicating that individuals tend to give socially desirable answers in reading surveys (Stanovich & West, 1989). This pattern emphasizes the necessity of incorporating objective measures of print exposure, such as the APT, to supplement subjective, self-reported measures, in order to capture more accurately the variation in individuals' reading experience.

Our approach holds significant advantages in terms of objectivity and efficiency. By harnessing the power of distributional semantics, we bypass the need for human categorization and rating, reducing the time and resources required for data processing. Additionally, it mitigates potential inaccuracies inherent in self-reported data, a problem particularly prevalent among younger adults, as shown in the varied correlations between objective and subjective preferences across genres among younger frequent and infrequent readers.

While the findings of this study are encouraging, we acknowledge that our approach has limitations. In the current form, the effectiveness of the APT and the measurement of print exposure relies heavily on the number of authors produced. Hence, our methodology may be more accurate in estimating the type of print exposure among experienced or frequent readers. For individuals with lesser reading exposure, the representativeness and accuracy of the participant vector might be compromised.

Looking forward, an intriguing avenue for future research would be exploring how to utilize the information derived from our approach, such as the cosine similarity between participant and genre vectors or the participant vectors themselves, to construct curated corpora that better explain individuals' behavioral data (e.g., lexical decision). By achieving this, we could potentially develop personalized models that more accurately capture the connection between individual lexical behavior and language exposure, thereby contributing to a richer understanding of psycholinguistic phenomena. In conclusion, our study presents a novel and promising approach to measuring the type of print exposure at the individual level. It not only enhances our understanding of the relationship between amount and type of print exposure across the lifespan, but also provides an efficient tool to capture these complex interactions objectively.

## 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.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly*, 285–303. Retrieved from `http://www.jstor.org/stable/748043`

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

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., ... Kurzweil, R. (2018, November). Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174). Brussels, Belgium: Association for Computational Linguistics.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In *Psychology of learning and motivation* (Vol. 22, pp. 193–225). Elsevier.

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, *73*(6), 841–855.

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science*, *42*(4), 1360–1374.

Johns, B. T., & Jamieson, R. K. (2019). The influence of place and time on lexical behavior: A distributional analysis. *Behavior Research Methods*, 1–16.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, *26*(1), 103–126.

Mata, R., & von Helversen, B. (2015). Search and the aging mind: The promise and limits of the cognitive control hypothesis of age differences in search. *Topics in Cognitive Science*, *7*(3), 416–427.

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*, 1–12.

McGeown, S. P., Duncan, L. G., Griffiths, Y. M., & Stothard, S. E. (2015). Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. *Reading and Writing*, *28*(4), 545–569.

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*(2), 267.

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.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Qiu, M., & Johns, B. T. (2020). Semantic diversity in paired-associate learning: Further evidence for the information accumulation perspective of cognitive aging. *Psychonomic Bulletin & Review*, *27*(1), 114–121.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379.

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

Struyf, A., Hubert, M., & Rousseeuw, P. (1997). Clustering in an object-oriented environment. *Journal of Statistical Software*, *1*(4), 1–30.

Taler, V., Johns, B. T., & Jones, M. N. (2020). A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the Canadian Longitudinal Study on Aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *75*(9), e221-e230.

Twenge, J. M., Martin, G. N., & Spitzberg, B. H. (2019). Trends in us adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print. *Psychology of Popular Media Culture*, *8*(4), 329–345.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. Retrieved from `http://jmlr.org/papers/v9/vandermaaten08a.html`

Wan, M., Misra, R., Nakashole, N., & McAuley, J. J. (2019). Fine-grained spoiler detection from large-scale review corpora. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, volume 1: Long papers* (pp. 2605–2610). Association for Computational Linguistics.

Wulff, D. U., De Deyne, S., Jones, M. N., Mata, R., & Aging Lexicon Consortium. (2019). New perspectives on the aging lexicon. *Trends in Cognitive Sciences*, *23*(8), 686–698.