Bilingualism: Language and Cognition

cambridge.org/bil

Intersample variance of second-language readers should not be overlooked

Victor Kuperman[1]

[1]McMaster University Hamilton, Ontario, Canada

Research Article

Cite this article: Kuperman, V. (2024). Intersample variance of second-language readers should not be overlooked. *Bilingualism: Language and Cognition*, 1–7 https://doi.org/10.1017/S1366728924000932

Received: 07 February 2024
Revised: 10 October 2024
Accepted: 12 October 2024

Keywords:
second language reading; reading comprehension; methodology; sampling

Corresponding author:
Victor Kuperman;
Email: vickup@mcmaster.ca

Abstract

Much of the literature on first (L1) second language (L2) reading agrees that there are noticeable behavioral differences between L1 and L2 readers of a given language, as well as between L2 speakers with different L1 backgrounds (Finnish vs German readers of English). Yet, this literature often overlooks potential variability between multiple samples of speakers of the *same* L1. This study examines this intersample variance using reading data from the ENglish Reading Online (ENRO) database of English reading behavior comprising 27 university student samples from 15 distinct L1 backgrounds. We found that the intersample variance within L2 readers of English with the same L1 background (e.g., two samples of Russian speakers) often overshadowed the difference between samples of L2 readers with different L1 backgrounds (Russian vs Chinese speakers of English). We discuss these and other problematic methodological implications of representing each L1 background with a single participant sample.

Highlights

- Studies often compare L1 and L2 readers or L2 readers with different L1 backgrounds
- Most studies use a single sample to represent each L1 background
- This practice overlooks the variance between samples from the same L1 background
- We show that this variance is sizable, even among advanced L2 readers of English
- Using multiple samples per L1 background is a recommended practice

1. Introduction

Meta-analyses of second-language reading point to several robust sources of variance in both reading comprehension and fluency (e.g., Bernhardt, 2011; Jeon & Yamashita, 2014, 2022; Melby-Lervåg & Lervåg, 2014). These sources include the reader's reading and speaking proficiency in their first (L1) language; proficiency in component skills in their second (L2) language; language and script distance between L1 and L2; as well as multiple extra-linguistic components (e.g., motivation, world knowledge and executive skills). Yet there is another source of variance in L2 reading behavior exists that is under-researched, namely, variance between samples of speakers that share an L1 and selected demographic and educational characteristics (Siegelman et al., 2023). A question that is rarely asked is this: Are group differences in L2 reading performance smaller when we compare participant samples with the same L1 than the differences seen in the samples of participants with different L1s. This paper discusses the implications that this question has for the practice and theory of second language research. It also provides an answer using a large-scale database of L1 and L2 reading behavior in English.

A common practice of a study in the field of second language reading is to recruit a single sample of participants for each first language. For instance, one sample each of Finnish and German L2 readers of English as well as one sample of Canadian L1 readers of English (e.g., Nisbet et al., 2022). In such studies, behavioral differences in English reading performance are– at least partly– attributed to differences in the L1 background of the participant groups. This conjecture relies on the extremely well-supported theoretical notion of cross-linguistic transfer, i.e., the effect that linguistic features of a person's L1 have on learning and reading proficiency in additional languages (e.g., Jarvis & Pavlenko, 2008; Marian & Kaushanskaya, 2007; Odlin, 2003). Since speakers of different L1s transfer a different set of features onto their reading behavior in L2, each distinct L1 background can be argued to have a characteristic behavioral signature (Nagata & Whittaker, 2013; Berzak et al., 2014). Cross-linguistic transfer can also explain the difference between L1 and L2 reading in a given language. L1 reading may be influenced by the cross-linguistic transfer from additional languages that the person speaks, but the transfer is rarely of the same magnitude as the L1 transfer onto L2. The notion of a distinct behavioral signature for every L1 finds support in recent computational models that are relatively accurate in telling apart L1 from L2 reading of English and identifying the specific L1 background of the L2 readers of English (e.g., Berzak et al., 2017; Skerath et al., 2023, see also Reich et al., 2022).

CAMBRIDGE UNIVERSITY PRESS

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press

We argue that the practice of offering a single participant sample per L1 background is potentially problematic for the following reason. It conflates the true effect of L1 background with the effects stemming from the characteristics of the specific sample. Consider the case of convenience pools of undergraduate university students, which are often recruited from studies of healthy adults' reading (Wild et al., 2022). L2 reading behavior of two groups of students drawn from two different universities may differ even if both groups consist of L1 speakers of the same language and were born and educated in the same country of testing. The difference may stem from multiple factors, including admission and attrition criteria related to L1 and especially L2 proficiency of respective universities; socio-economic and linguistic composition of the student body; the amount of explicit instruction or practice in the L2 through university curricula; the universities' relative prestige and competitiveness; the networking and collaborative opportunities, including international student exchanges and internships; and others. An even more important factor may be the cross-university differences in L1 literacy of respective students. In line with the interactive compensatory model of reading acquisition (Stanovich, 1980), Bernhardt (2011) estimated the contribution of L1 proficiency to be 20% of the variance in L2 reading comprehension. It is logically possible then, that the intersample within-L1 differences would be comparable or larger than the between-L1 differences. If confirmed, this possibility would put a question mark over just how distinct the L1-driven behavioral signatures are in L2 reading behavior. It would also constrain the ability of single-sample studies to generalize their conclusions regarding behavioral effects caused by the L1 background beyond the specific samples they consider.

One of the very few studies that have the data to address these questions is the ENglish Reading Online (ENRO) database of L1 and L2 English reading and component skills of English reading (Siegelman et al., 2023). This database represents multiple samples of university students, some of which share the same L1 and were drawn from the same country, and others which do not, see details below. Siegleman et al. quantified the contribution of the within-L1 intersample variability in English reading comprehension and fluency and listening comprehension. This variability accounts for about 1.5% of variance in reading and listening comprehension and 3.1% in reading rate. In all measures, this amount of explained variance is on par with the amount of variance explained by the L1–L2 distinction. This finding suggests that the often-omitted source of variance– that between samples of same-L1 speakers– is worth examination.

The ENglish Reading Online (ENRO) project makes available data from 36 participant samples of L1 and L2 speakers of English collected in 19 countries. It represents relatively advanced readers of English as L2. Most participating universities in countries where English is not a dominant language have an entrance requirement of a certain level of English proficiency and use instructional materials in English (see Siegelman et al., 2023 for discussion). Importantly, for our purposes, Russian, Italian and German as L1s are represented by two or three university samples each; English is the L1 of 9 participant samples and 10 additional samples represent one L1 background each. We estimate behavioral similarities and differences within participant groups that share the same L1 background and between groups that differ in their L1 background. We base our comparisons on a battery of 12 indices tapping both into the reading performance and component skills of reading in English. These include reading fluency (reading rate in words per minute) and comprehension accuracy; tests of listening comprehension, grammatical, orthographic and vocabulary knowledge; decoding and lexical decision. Due to the practical impossibility of locating and conducting comparable tests of component skills proficiency in 16 distinct languages, the ENRO offers no data on L1 literacy for non-English-dominant participants. We come back to this point in the Limitations section below.

If the present data were to support the notion of the L1-specific behavioral signature, we expect differences between participant groups with the same L1 background (but from different universities) to be smaller than those between participant groups with dissimilar L1s. Finding evidence to the contrary would mean that the common practice of representing each language with a single sample may lead to a biased interpretation of experimental results.

2. Method

This study reanalyzes data from the multi-lab ENRO project. The data file is openly available in the Open Science Framework at https://osf.io/epyu8, and supplementary information at https://osf.io/gzyqf/. Full details on the methodology, recruitment method, participant samples and the test battery are available in Siegelman et al. (2023).

2.1. Participants

The ENRO database includes data on English text reading and component skills of English proficiency from 7338 participants, recruited from 30 partner sites. The coverage of the ENRO data is 19 countries and 16 source languages. (The source language is defined as the language of instruction in the university where data was collected.) Twenty-eight of the samples were recruited via university participant pools, while the remaining two were recruited online using the crowdsourcing platform Prolific. Nine samples had English as the source language (4 in Canada, 1 in New Zealand, 1 in the UK, 2 in the USA and 1 sample of L1 English-speaking participants from the US, UK and Canada recruited through Prolific). These samples included both L1 and non-L1 speakers of English. The remaining 21 samples were collected in the countries and universities where English is a non-dominant language. These samples included participants with the same first language as the language of instruction in the university. Since we cannot ascertain whether English is the second, third or additional language of participants in these samples, for simplicity we designated these samples as L2 speakers of English. Importantly, the data pool contains three samples collected in Germany, three in Russia and two in Italy. These replicated samples are at the core of our comparative analyses.

Several samples were excluded from consideration. Thus, we excluded all non-L1 speakers of English who were recruited through universities with English as the language of instruction. The reason is that L2 participant groups in English-speaking universities are mixed in terms of their first language. We also excluded two participant samples– L1 English speakers and L1 Dutch speakers– recruited via the Prolific platform. These samples represent a mixture of universities and thus are not usable for the present purposes. A final sample that we excluded were L1 speakers of Slovenian (University of Ljubljana, Slovenia). Due to a technical error, participants in this sample did not complete the listening comprehension test, and thus they cannot be compared to other samples on all tests. The remaining data pool consisted of 6042 participants and 27 participant samples representing 15 first

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press

Table 1. Information regarding participants in the sample units

Unit	Country	University	English status	Source language	N
ca_mcg_english	Canada	McGill U	L1	English	61
ca_mcm_english	Canada	McMaster U	L1	English	1895
ca_ua_english	Canada	U of Alberta	L1	English	271
ca_uo_english	Canada	U of Ottawa	L1	English	759
nz_uv_english	New Zealand	Victoria U of Wellington	L1	English	120
uk_uos_english	UK	U of Southampton	L1	English	122
usa_csi_english	USA	CUNY	L1	English	179
usa_msu_english	USA	Michigan State U	L1	English	147
ar_utdt_spanish	Argentina	U Torcuato Di Tella	L2	Spanish	102
be_ugh_dutch	Belgium	U Ghent	L2	Dutch	205
be_ulb_french	Belgium	U Libre de Bruxelles	L2	French	105
de_du_german	Germany	Heinrich-Heine-U Düsseldorf	L2	German	53
de_gu_german	Germany	U Goettingen	L2	German	146
de_ku_german	Germany	Katholische U Eichstatt-Ingolstadt	L2	German	104
il_huji_he_hebrew	Israel	Hebrew U	L2	Hebrew	112
il_huji_ar_arabic	Israel	Hebrew U	L2	Arabic	101
in_iitk_hindi	India	Indian Institute of Tech, Kanpur	L2	Hindi	157
it_si_italian	Italy	SISSA	L2	Italian	151
it_unimib_italian	Italy	U of Milano-Bicocca	L2	Italian	221
jp_nu_japanese	Japan	Nagoya U	L2	Japanese	129
mn_kho_mongolian	Mongolia	Khovd State U	L2	Mongolian	51
ru_hse_russian	Russia	HSE Moscow	L2	Russian	73
ru_spb_russian	Russia	St Petersburg U	L2	Russian	59
ru_tu_russian	Russia	Tomsk U	L2	Russian	164
rs_bg_serbian	Serbia	U of Belgrade	L2	Serbian	301
th_tu_thai	Thailand	Thammasat U	L2	Thai	101
tw_ntnu_chinese	Taiwan	National Taiwan Normal U	L2	Chinese	153

Notes: *U* = University; *USA* = United States of America; *UK* = United Kingdom; *CUNY* = City University of New York College of Staten Island.

languages. Table 1 summarizes information about these samples. See Siegelman et al. (2023) for information about mean age, gender and education, details regarding compensation and ethics clearance.

2.2. Materials, procedure and apparatus

The ENRO battery of assessments in English included tests of reading comprehension, tests of component skills of reading and demographic and language background questionnaires. All participants completed all assessments using the battery. The full list of assessments (including references) and their estimated reliability are available in Siegelman et al. (2023). Below, we only briefly describe the tests that the paper uses in the analyses.

The test of reading comprehension consists of 15 texts followed by three 4-alternative-forced-choice comprehension questions designed to test individuals' factual understanding and inferential ability. Texts are based on training materials for the ACCUPLACER Reading test and the English as Second Language Reading Skills test (https://accuplacer.collegeboard.org/students/prepare-for-accuplacer/practice). This test has two outcome variables: (i) percent correct out 45 comprehension questions and (ii) reading rate measured in words per minute. The listening comprehension test consists of 5 recorded texts presented to participants for audition, excerpted from Sommers et al. (2011). Each text is followed by five 4-alternative-forced-choice comprehension questions. The outcome variable of this test is (iii) the percent of questions answered correctly out of the 25 questions. Additional tests tap into component skills of English reading proficiency. The grammaticality judgment task as well as tests of spelling recognition (adapted from Andrews & Hersch, 2010), vocabulary knowledge (adapted from Nation & Beglar, 2007), orthographic awareness (Siegel et al., 1995) and word-chain text segmentation each produced a score based on percent correct of responses: These were outcome variables iv-viii. Two additional tests are the Lexical Test for Advanced Learners of English (LexTALE, Lemhöfer & Boersma, 2012) and the lexical decision task. Both these tests produce the accuracy (percent correct) and response time as outcomes (ix-xii). These 12 outcome variables are used to compare the performance between participant groups. The ENRO tasks were administered online, using a custom-tailored web-based data collection platform. The entire study typically took about 1.5 hours to complete.

2.3. Statistical considerations

The central statistical decision that this and similar studies faces is how to quantify differences between groups of participants, i.e., what metric of an intersample distance to consider. One facet of this decision is what behavioral outcomes to choose for calculating the distance. The options include (1) considering measures of reading performance only (i.e., reading rate and comprehension accuracy in the present data and eye-movement measures in research cited above) or (2) a combination of the reading measures and component skills of reading, including listening comprehension, vocabulary knowledge, decoding skill and many others.

Another facet to decide on is how to aggregate the individual's performance in each group to enable group comparisons. One possibility (a) is to calculate the group mean in each behavioral measure and represent the performance of every group as a multidimensional vector of the mean scores. The distance between the vectors representing each pair of groups is then used as a measure of separation between those groups. To give each behavioral measure an equal weight, the group means for each measure are normalized before calculating the Euclidian distance between pairs of vectors.

An alternative way (b) of quantifying the pairwise distance between groups is to consider each behavioral measure separately and calculate the overlap in the distributions of individual values for this measure between the two groups (Kuperman et al., 2024). The overlap between the group-specific distributions can be quantified as a distribution-free overlapping index provided in the overlapping package in R (Pastore & Calcagni, 2019). Here, the distribution of values in a group is first calculated using the Gaussian-density estimation. The similarity between the values in each group is then computed as the overlap between these distributions and expressed as the overlap index OV. The index stands for the percentage of the area of the density distribution that is shared between groups and ranges from 0 to 1. For example, an OV index of .95 indicates that 95% of the language-specific distributions are overlapping. The similarity between any two groups can be estimated as the mean of the overlap indices for those groups across all behavioral measures. The distance between a pair of groups is calculated as 1 minus the mean overlap index.

Once the distance matrix for all pairwise group comparisons is obtained, we can compare distances between groups of L1 vs L2 readers of English, as well as distances between multiple groups of speakers sharing the same L1 vs groups of speakers with different L1s. We opted for the approach that combines options (2) and (b). Namely, we considered all behavioral measures in the ENRO data (both reading performance and component skills) and used the overlap metric for calculating pairwise group distances. The inclusion of component skills along with measures of reading into the estimation of the group distance reflects the importance of those skills for acquisition of L2 reading (e.g., Jeon & Yamashita, 2022; Melby-Lervåg & Lervåg, 2014). Also, we preferred using the overlap index over the mean values because the former represents the entire distributions rather than the pointwise estimate of central tendency per group. Critically, basing our analyses below on all other available options (1a, 1b, 2a) produced results that are nearly identical to the statistical outcomes of 2b. Thus, our findings below are stable even in view of the researcher's degrees of freedom.

All analyses below were made using the statistical platform R v 4.2.2 (R Core Team, 2022). Function cmdscale() was used to apply multidimensional scaling to visualize the distances and ggplot2 library (Wickham, 2011) to plot the results.

3. Results

We analyzed the pool consisting of 6042 participants that represent 27 university-based samples and 15 distinct L1s. Twelve behavioral outcomes recorded in the ENRO database were used for the group comparison between samples. As described above, the overlap between each pair of groups was calculated separately for every behavioral measure, yielding 12 OV indices for the pairwise comparison. The distance between each pair of groups was calculated as one minus the mean of the OV indices across all behavioral measures. The distance ranged from 0 to 1. Iterating this estimation across all group pairs yielded a 27 × 27 distance matrix. For visualization, we applied classic multidimensional scaling to the distance matrix and mapped the groups onto the two-dimensional scatter plot, see Jaworska and Chupetlovska-Anastasova (2009) for a review of multidimensional scaling. Figure 1 provides the plot and additionally highlights distances between non-English-dominant participant groups representing the same L1. All analyses below are based on the original distance matrix rather than the outcomes of the multidimensional scaling. Because multidimensional scaling reduces the number of dimensions to depict distances between groups, it adds distortion to the original distance estimates.

Figure 1 illustrates several interesting points regarding intersample similarities and differences. First, participant groups with English as L1 (shown in red) cluster much closer to each other than do groups with English as L2 (shown in green). Distances between L1 groups are significantly smaller than between L2 groups ($\beta = 0.28$, $SE = 0.07$, $p < 0.001$), and both are significantly smaller than the distances between L1 and L2 groups ($\beta = 0.73$, $SE = 0.07$, $p < 0.001$ relative to L1 as the reference level).

The key question of this paper is how the within-L1 intersample differences compare to the intersample differences between L1s. One approach to this question is to determine whether two groups (A and B) that share an L1 background are also each other's nearest neighbors. i.e., whether the shortest distance from group A to another group is the A-to-B distance, and vice versa. (We made this determination based on the 27 × 27 matrix of distances as defined in the Method rather than the multidimensional scaling representation of the distances.) The participant groups that share English as their L1 background provide a very clear answer. For each L1 English group (marked in red in Figure 1), their nearest neighbor is another group with English as L1. The only exception is the CUNY Staten Island (USA) sample, which has as its nearest neighbor another sample of participants highly fluent in English (IITK, India). This finding implies a high consistency in English reading behavior and component skills among L1 speakers of English. Thus, behavioral outcomes shown by virtually all samples of university-level L1 English speakers in our data can be generalized to university-level L1 English speakers across English-dominant countries and universities.

We now turn to the non-English-dominant participant samples. Figure 1 marks in color eight same-language distances, i.e., the pairwise distances between three groups of L1 German speakers, three groups of L1 Russian speakers and two groups of L1 Italian

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press

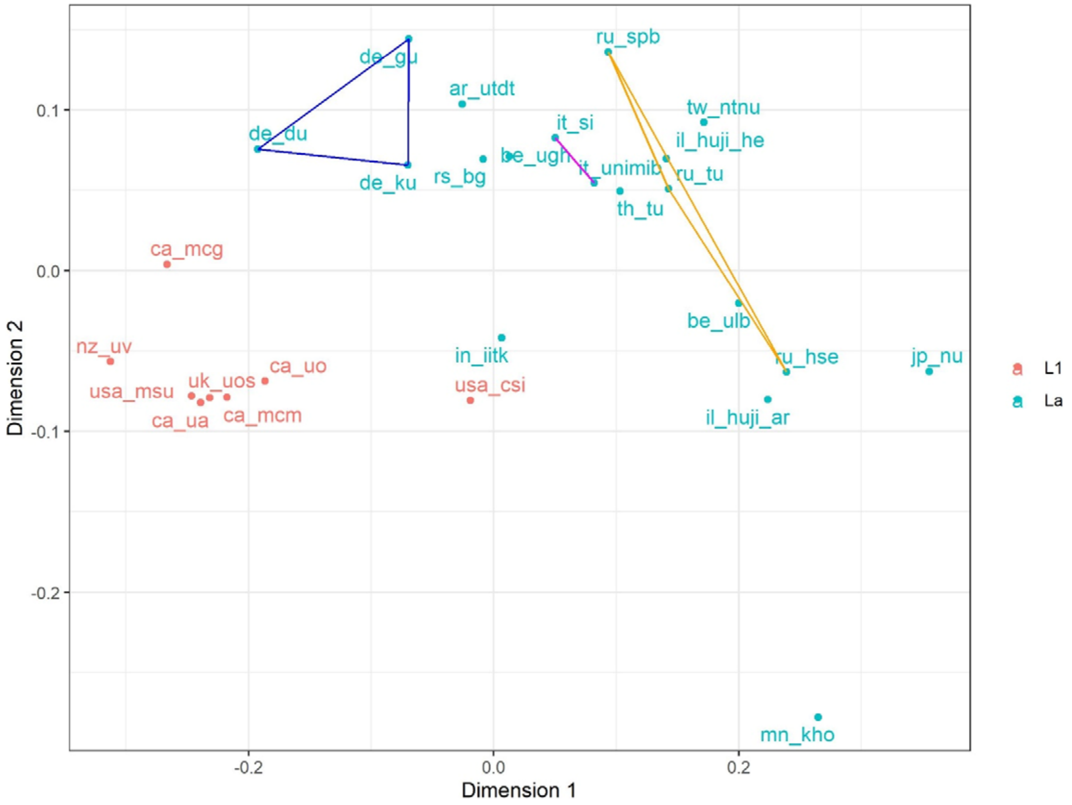

Figure 1. Multidimensional scaling of distances between ENRO participant groups. Groups of L1 speakers of English are shown in red and those of L2 in green. Distances between same-L1 groups are shown in blue (German), orange (Russian) and purple (Italian).

Table 2. Ranks of distances between groups of participants with the same L1. For sample codes, see Table 1.

From/to German	Rank distance	From/to Russian	Rank distance	From/to Italian	Rank distance
de_du/de_gu	2	ru_hse/ru_spb	12	it_si/it_unimib	1
de_du/de_ku	1	ru_hse/ru_tu	3	it_unimib/it_si	1
de_ku/de_du	7	ru_spb/ru_hse	14		
de_ku/de_gu	1	ru_spb/ru_tu	1		
de_gu/de_du	5	ru_tu/ru_hse	8		
de_gu/de_ku	1	ru_tu/ru_spb	3		

speakers. Visually, some of these groups are close (e.g., de_gu and de_ku; it_si and it_unimib; and ru_tu and ru_spb), while others are at a considerable distance from their counterparts (de_du; ru_hse). As a first step, we applied the non-parametric one-tailed Wilcoxon rank sum test to determine whether the same-L1 distances are shorter on the average than between-L1 distances: They are significantly so ($W = 237$, $p = 0.004$).

The second step applies a more stringent test: We ask whether the same-L1 groups of participants are the nearest neighbors to one another, i.e., have the shortest pairwise distance compared to all other groups. For every group that has the same-L1 counterpart, we estimated the rank of the distance between that group and each counterpart to a total of 14 group pairs. With 27 groups under comparison, the average rank distance is 13. The rank distance of 1 from group A to group B implies that group B is the nearest neighbor of group A. As is clear from Figure 1, this does not guarantee that group A is also the nearest neighbor for group B. While the A-to-B distance is the same as the B-to-A distance, groups A and B are at different distances from other neighbors. Table 2 summarizes rank distances for 14 same-L1 group pairs.

Data in Table 2 reveal that some pairs of the same-L1 participant groups are indeed each other's nearest neighbors in the multidimensional space that represents distributions of behavioral outcomes of all groups. This is true of the two available samples of L1 Italian speakers and of the L1 German speakers from University of Duesseldorf (de_du) and their distance from the other two samples from Germany. Finally, the two samples of German speakers from University of Goettingen (de_gu) and from University of Keiserslautern, de_ku) are each other's nearest neighbors[1].

Yet the above pattern was far from universal. Seven out of 14 samples in Table 1 have at least one different-L1 participant

[1] Because of the distortion added by multidimensional scaling, Figure 1 shows the Argentine group (ar_utdt) rather than de_ku as the closest neighbor of de_gu, even though the distance between de_gu and ar_utdt is slightly larger than that between de_gu and de_ku (0.129 vs 0.123).

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press

group at a shorter distance to them than the group of participants that speaks the same L1. The number of "intervening" different-L1 groups ranges from 2 (distance from ru_tu to ru_spb) to 13 (from ru_spb to ru_hse). Across all groups presented in Table 2, the median number of different-L1 groups that are more similar to a given group than its same-L1 counterpart is 1.5 and the mean number is 3.2 (SD = 3.9). We return to this finding in the General Discussion.

4. General discussion

The field of second language reading has a major focus on the role of the reader's first language (e.g., Droop & Verhoeven, 2003; Jiang, 2011; Upton & Lee-Thompson, 2001). It is customary for experimental studies in this field to (partly) attribute behavioral differences between L1 and L2 readers or between L2 readers with different L1s to the specific linguistic features of their respective L1s. Yet many such studies recruit a single participant sample to represent each given L1 background. Therefore, such attribution may conflate the true effect of the L1 background (which the participant sample shares with all speakers of that L1) and the characteristics of the specific sample. In studies that recruit healthy adult participants from convenience pools of university students, this practice amounts to selecting students from a specific program or university to be representative of students from other programs and universities in this country speaking the same L1. Yet, as we discuss in the Introduction, universities may systematically vary in the language proficiency of their students, especially when it comes to their L2 proficiency. This paper offers an in-depth examination of the source of variance related to the intersample variability within L1. We tested whether within-L1 group differences could overshadow the behavioral signature that each L1 background generates in the L2 reading behavior via cross-sectional transfer.

Siegelman et al.'s (2023) analysis of the ENRO database demonstrates that intersample variability within-L1 and within-country accounts for a nontrivial amount of explained variance in English reading comprehension and fluency. The present study made use of 27 groups of university students from the same dataset. Nine of these samples have English as their L1. Among the remaining 18 groups of L2 speakers of English, 10 represented single samples of a specific L1 background, and eight had at least one more counterpart with the same L1 (three samples of German, three of Russian and two of Italian), and 10 additional groups represent one L1 each. We estimated pairwise distances between all groups using as a metric the overlap in the distributions of their test scores. Based on 12 tests of reading comprehension, reading fluency and multiple component skills of English proficiency, these distances enabled a comparison of both within-L1 and between-L1 variability.

Our findings are quite straightforward, see Figure 1. First, L1 English participant groups are distinguishable in their behavior from L2 English participant groups. The distances between L1 English and L2 samples were on average much larger than within the set of L1 English samples. This suggests that a classifier that seeks to determine whether a given participant or a group belong to L1 or L2 readers would be highly successful. Second, L1 English participant groups are more homogeneous than L2 groups, such that the intersample distances within the former set were significantly shorter than within the latter set. Taken together, these findings imply that an L1 English group cannot be easily confused in its behavior with a group of L2 English speakers. They also imply that a single student sample of L1 English speakers can be considered representative of a large variety of student samples of L1 English speakers from multiple universities and English-dominant countries (four countries and eight universities in the ENRO data).

The third finding concerns the L2 readers of English and relates directly to the goal of our study. We found that the shared L1 background and country (and demographic and educational characteristics of university student samples) do not guarantee high similarity in L2 behavior. It is true that, on average, participant groups sharing an L1 are closer to one another than the groups with different L1 in terms of their L2 reading and component skills. This converges well with the notion of the cross-linguistic transfer that leads to the L1 background producing a characteristic behavioral signature. Yet, in half of the pairwise comparisons (Table 2), a participant group that has the same-L1 counterpart performed more similarly to one or more different-L1 groups than to that counterpart. To take an extreme example, two samples of L1 Russian readers of English (ru_hse and ru_spb) were as far from one another behaviorally as any two randomly drawn samples of L2 readers of English in our data. It is worth mentioning that both universities from which these samples originated are located in the two largest cities of Russia (Moscow and St Petersburg, respectively), are highly selective and are highly ranked in their country (ranks 4 and 2 in the Shanghai ranking, respectively). Thus, very substantial within-L1 differences can be observed even across institutions of comparable standing. On average, 3.2 different-L1 samples proved to be closer to the samples under comparison than the same-L1 samples (median = 1.5).

These findings suggest that the behavioral signature that the L1 background yields in L2 reading and component skills is present but not sufficiently distinct. The ability of researchers to draw conclusions regarding the specific effects of the L1 background (or their similarity/difference from other L1s or L2s) is limited. Specifically, it is contingent on the specific university (or a program within university) in which participant recruitment takes place. We cannot presently estimate the exact extent of the systematic error stemming from the practice of drawing a single participant sample per L1 and glossing over the within-L1 intersample variability. Still, the implicit assumption that a single sample of participants is representative of all speakers of the given L1 (even within highly proficient student populations) cannot be taken for granted. Validity of this assumption is dependent on multiple linguistic, educational and social factors, many of which require extensive further study. Our recommendation for studies of second-language reading is to provide multiple samples of participants for each first-language background. Only then can researchers tease apart the effect of L1 from the specifics of their samples.

4.1. Limitations and future directions

This study only lists but does not explore the reasons that may underline inter-university differences in (L1 or L2) English proficiency within a country. One such reason may be the differences in L1 literacy, which contribute as much as 20% of explained variance in L2 reading comprehension (Bernhardt, 2011). ENRO as the data source only incorporates tests of English proficiency and does not tap into non-English L1 literacy comprehension or its component skills.

Most likely, other reasons involve the interaction of federal or local funding for language education; language requirements of secondary and tertiary schools for enrollment and degree attainment; university standing; prestige and popularity of the L2 as a language of professional, social or personal communication and a

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press

myriad of other factors. Presently, we are unable to pin down specific sources of interuniversity variability and relegate this intriguing question to future research.

The methodological implications of the present findings are quite straightforward. The common practice of using a sole sample of L2 readers from a given L1 background runs the demonstrable risk of producing biased data, where the linguistic properties of that L1 are conflated with the specific characteristics of the participant sample. Our data suggest that the issue is not drastic if the goal is to compare the performance of L1 readers of the language with L2 readers of that language with a specific background (e.g., readers of English with the English vs Chinese L1 background). The separation in the performance of L1 vs L2 readers of English was nearly categorical. The issue gains importance when comparing reading and related skills among participant groups with different L1 backgrounds (e.g., Chinese- vs Spanish-speaking readers of English as L2). Accuracy of such comparisons demonstrably depends on the (university) populations from which the speakers are sampled. Same-L1 groups were as likely to be each other's nearest neighbors in reading performance as the nearest neighbors of a different-L1 group. We recommend that studies pursuing comparisons of different L1 backgrounds recruit multiple samples of readers from each background, each from a comparable but different origin. This practice would allow disentangling the effect of the L1 background from the effect of the sample.

A more global future direction may be to evaluate the variability of L2 reading proficiency across educational institutions within a country, since this knowledge would give the best indication of how representative a participant sample is in the given experimental study. While the ENRO data is relatively unique in offering two or three samples of several non-English language backgrounds, it is still insufficient both in the number of samples and coverage for a comprehensive study of intersample variability. It would be beneficial to develop behavioral mega studies in which more (and ideally all) languages are represented by multiple participant groups and the intersample variability is estimable. Multi-lab collaborations are a promising way to obtain the necessary but currently sparse experimental and mega-study data.

Acknowledgments. The author's contribution was supported by the Social Sciences and Humanities Research Council of Canada Partnered Research Training Grant (895-2016-1008; Libben PI), Insight Grant (435-2021-0657; Kuperman, PI) and the Canada Research Chair (Tier 2; Kuperman, PI). Thanks are due to Daniel Schmidtke and Jordan Gallant for insightful comments on this manuscript.

References

Andrews, S., & **Hersch, J.** (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, **139**(2), 299–318.

Bernhardt, E. B. (2011). *Understanding advanced second-language reading*. Routledge.

Berzak, Y., **Nakamura, C.**, **Flynn, S.**, & **Katz, B.** (2017). Predicting native language from gaze. *arXiv preprint arXiv:1704.07398*.

Berzak, Y., **Reichart, R.**, & **Katz, B.** (2014). Reconstructing native language typology from foreign language usage. *arXiv preprint arXiv:1404.6312*.

Droop, M., & **Verhoeven, L.** (2003). Language proficiency and reading ability in first-and second-language learners. *Reading Research Quarterly*, **38**(1), 78–103.

Jarvis, S., & **Pavlenko, A.** (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Jaworska, N., & **Chupetlovska-Anastasova, A.** (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, **5**(1), 1–10.

Jeon, E. H., & **Yamashita, J.** (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, **64**, 160–212.

Jeon, E. H., & **Yamashita, J.** (2022). L2 reading comprehension and its correlates. E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp.29–86).

Jiang, X. (2011). The role of first language literacy and second language proficiency in second language reading comprehension. *The Reading Matrix*, **11**(2).

Marian, V., & **Kaushanskaya, M.** (2007). Cross-linguistic transfer and borrowing in bilinguals. *Applied Psycholinguistics*, **28**(2), 369–390.

Melby-Lervåg, M., & **Lervåg, A.** (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first-and second-language learners. *Psychological bulletin*, **140**(2), 409–433.

Nagata, R., & **Whittaker, E.** (2013, August). Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 1137–1147).

Nisbet, K., **Bertram, R.**, **Erlinghagen, C.**, **Pieczykolan, A.**, & **Kuperman, V.** (2022). Quantifying the difference in reading fluency between L1 and L2 readers of English. *Studies in Second Language Acquisition*, **44**(2), 407–434.

Odlin, T. (2003). Cross-linguistic influence. In *The handbook of second language acquisition* (pp. 436–486).

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Reich, D. R., **Prasse, P.**, **Tschirner, C.**, **Haller, P.**, **Goldhammer, F.**, & **Jäger, L. A.** (2022, June). Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on eye tracking research and applications* (pp. 1–8).

Siegelman, N., **Elgort, I.**, **Brysbaert, M.**, **Agrawal, N.**, **Amenta, S.**, **Arsenijević Mijalković, J.**, … & **Kuperman, V.** (2023). Rethinking first language–second language similarities and differences in English proficiency: Insights from the ENglish Reading Online (ENRO) project. *Language Learning*, **74***(1)*, 249–294.

Skerath, L., **Toborek, P.**, **Zielińska, A.**, **Barrett, M.**, & **Van Der Goot, R.** (2023, July). Native language prediction from Gaze: A reproducibility study. In *Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 4: Student research workshop)* (pp. 152–159).

Sommers, M. S., **Hale, S.**, **Myerson, J.**, **Rose, N.**, **Tye-Murray, N.**, & **Spehar, B.** (2011). Listening comprehension across the adult lifespan. *Ear and hearing*, **32**(6), 775–781.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, **16***(1)*, 32–71.

Siegel, L. S., **Share, D.**, & **Geva, E.** (1995). Evidence for superior orthographic skills in dyslexics. *Psychological science*, **6**(4), 250–254.

Upton, T. A., & **Lee-Thompson, L. C.** (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition*, **23**(4), 469–495.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, **3**(2), 180–185.

Wild, H., **Kyröläinen, A. J.**, & **Kuperman, V.** (2022). How representative are student convenience samples? A study of literacy and numeracy skills in 32 countries. *Plos One*, **17**(7), e0271191.

https://doi.org/10.1017/S1366728924000932 Published online by Cambridge University Press