Research Article

Geographic variability in limited English proficiency: A cross-cultural study of cognitive profiles

Iulia Crişan[1] , Sami Ali[2], Laura Cutler[2], Alina Matei[1] , Luisa Avram[1] and Laszlo A Erdodi[2]

[1]Department of Psychology, West University of Timişoara, Timişoara, Romania and [2]Department of Psychology, University of Windsor, Windsor, Canada

Abstract

Objective: This study was designed to evaluate the effect of limited English proficiency (LEP) on neurocognitive profiles. **Method:** Romanian (LEP-RO; $n = 59$) and Arabic (LEP-AR; $n = 30$) native speakers were compared to Canadian native speakers of English (NSE; $n = 24$) on a strategically selected battery of neuropsychological tests. **Results:** As predicted, participants with LEP demonstrated significantly lower performance on tests with high verbal mediation relative to US norms and the NSE sample (large effects). In contrast, several tests with low verbal mediation were robust to LEP. However, clinically relevant deviations from this general pattern were observed. The level of English proficiency varied significantly within the LEP-RO and was associated with a predictable performance pattern on tests with high verbal mediation. **Conclusions:** The heterogeneity in cognitive profiles among individuals with LEP challenges the notion that LEP status is a unitary construct. The level of verbal mediation is an imperfect predictor of the performance of LEP examinees during neuropsychological testing. Several commonly used measures were identified that are robust to the deleterious effects of LEP. Administering tests in the examinee's native language may not be the optimal solution to contain the confounding effect of LEP in cognitive evaluations.

Keywords: bilingualism; cross-cultural comparison; neuropsychological testing; memory and learning tests; executive function; Boston naming test

(Received 15 August 2022; final revision 24 January 2023; accepted 18 April 2023; First Published online 29 May 2023)

Introduction

Most neurocognitive tests have been developed in North America and normed on native English speakers (NSEs). Normative systems typically focus on age, education, gender, or race (Abeare et al., 2019; Heaton et al., 2004, 2009) and tend to ignore variability in language proficiency (Gasquoine et al., 2007; Gasquoine, 1999). Limited English proficiency (LEP) refers to a continuum of deficits in phonology (systematic phoneme substitutions characteristic of foreign accents), lexicon (limited vocabulary and speed of retrieval), and syntax (deviation from grammatical rules) attributable to late-life language acquisition (i.e., outside the sensitive period) in the context of normal verbal skills in the individual's mother tongue. In other words, LEP is a *learned deficit* reflecting a delay in exposure to English.

Recent research demonstrated that LEP can be a significant confound in test result interpretation even in cognitively high-functioning examinees (Ali, Brantuo, et al., 2022; Erdodi et al., 2017a). Consequently, existing norms may not apply to individuals with LEP (Celik et al., 2020; Funes et al., 2016; Gasquoine & Gonzales, 2012), as they systematically underestimate *verbal skills in general* – while perhaps providing an accurate measure of English proficiency. As the world grows diverse due to migration and the percentage of bilinguals increases both in Europe and the USA (Eurostat, 2018; Ryan, 2013), so do the chances of encountering patients with LEP in clinical settings.

Therefore, understanding the impact of LEP on cognitive testing is of immediate practical interest.

Recent reviews (Antoniou, 2019; Celik et al., 2020) have outlined bilinguals' advantages and disadvantages in different cognitive tasks. The tasks' level of verbal mediation further complicates the interpretation of cognitive profiles associated with LEP. Verbal mediation has recently been referenced in LEP research to classify neuropsychological instruments based on *the extent to which intact language skills and/or native-level proficiency in the language of administration is required for the test to provide a valid measure of its target construct* (Brantuo et al., 2022). Throughout this paper, we refer to "*verbal*" tests as having *high verbal mediation,* indicating that verbal skills are central to optimal performance. In contrast, we refer to "*non-verbal*" tests (i.e., tasks designed to measure visual–perceptual skills; Gasquoine et al., 2007) as having *low verbal mediation*.

Studies comparing NSE and LEP groups on tests administered in English have yielded contradictory results (Boone et al., 2007; Gasquoine et al., 2007; Kisser et al., 2012). On the one hand, there are reports of NSE performing better on verbal but not nonverbal tasks, such as tests of visuospatial abilities (Boone et al., 2007; Kisser et al., 2012). On the other hand, significant language administration effects in Spanish-English bilingual groups were documented for some (e.g., letter fluency, Stroop Color and Word

trials) but not other verbal tasks (i.e., verbal learning, Digit Span; Gasquoine et al., 2007).

Theoretically, nonverbal tests should be immune to LEP. Indeed, NSE norms for certain visuospatial measures can be applied to Spanish-speaking LEP samples without increasing false-positive rates (Gasquoine & Gonzales, 2012; Gasquoine et al., 2007). Similarly, Walker et al. (2010) found no differences between NSE and LEP participants with different English proficiency levels on several tests (e.g., Digit Symbol, Matrix Reasoning). However, Funes et al. (2016) demonstrated that administering tests in English to Spanish-speaking participants may overestimate deficits even on non-verbal tasks (e.g., Digit-Symbol Coding, Block Design).

Predictably, NSE outperform LEP on multiple verbal tasks, including auditory attention (Digit Span/Word Span; Durand-Lopez, 2020; Mattys et al., 2017; Walker et al., 2010; Yoo & Kaushanskaya, 2012), executive functions (Stroop; Coderre et al., 2013; Singh & Mishra, 2013; Tse & Altarriba, 2012), object naming or verbal fluency (BNT-15; Ali, Elliott, et al., 2022; Brantuo et al., 2022; Erdodi et al., 2016). However, not all tests that, at face value, appear to have high verbal mediation are equally affected by LEP. Some (animal fluency, BNT, Complex Ideational Material, single word reading) are particularly sensitive to it, whereas others (speeded reading, Digit Span) seem surprisingly robust to LEP (Ali, Brantuo, et al., 2022; Ali, Elliott, et al., 2022; Brantuo et al., 2022; Kousaie et al., 2014; Papageorgiou et al., 2019). Additionally, the *degree* of LEP (Coderre et al., 2013; Marian et al., 2013; Roselli et al., 2002; Tse & Altarriba, 2012; Walker et al., 2010), task difficulty (Durand-Lopez, 2020), and even the examinees' mother tongues (Ardila, 2020; Mattys et al., 2017) can also mediate test performance. Such results re-iterate the fact that LEP is a heterogeneous category – treating it simply as the opposite of NSE may overlook important within-group trends that could further inform research on cross-cultural neuropsychology.

Such divergent findings raise several questions about the effect of English proficiency on neuropsychological testing: which cognitive tasks are most affected by LEP? Is the neurocognitive profile associated with LEP more complex than a predictable pattern of deficits based on the level of verbal mediation? Are there meaningful subtypes within LEP? This study was designed to provide tentative answers to these questions. Since most prior research on LEP has been based on US Spanish-English bilinguals, we recruited two geographically and linguistically diverse LEP samples to test the limits of generalizability.

These two bilingual samples (Arabic-dominant students from Canada and Romanian-dominant students from Romania) were recruited to examine the geographic, cultural, and linguistic variability in cognitive profiles associated with LEP. Their main shared commonality is their non-NSE status. In contrast, the differences between them are significant and multifactorial: different native languages (Romanian versus Arabic), writing systems (26-letter Latin alphabet versus the abjad), directions of writing/reading (left-to-right versus right-to-left), educational systems, the broader cultural context (Central Europe versus the Middle East) and cultural identity (Romanian versus Arabic Canadians), the relative homogeneity within the groups and immigration status (all Romanian participants were born and raised in Romania and recruited from a single university whereas the Arabic participants immigrated from various countries) could potentially influence performance on neuropsychological testing. Therefore, comparing the Romanian and Arabic samples provided a robust method for examining whether LEP should be considered a unitary or a heterogeneous construct.

We made the following predictions: (1) All participants with LEP would perform worse than NSEs and below the US normative mean on verbal tests; (2) There would be no difference between NSE and LEP on nonverbal tests; (3) Within participants with LEP, performance on verbal tests would differ as a function of the *relative* level of English proficiency.

Method

Participants

Data were collected from 113 cognitively healthy university students (98 women; $M_{Age} = 22.7$; $SD = 5.6$; $M_{Education} = 14.2$; $SD = 2.0$). Participants were recruited from two countries (the Western region of Romania and South-Central Canada) and divided into three samples: Romanian-English bilinguals with LEP ($n = 59$; LEP-RO), Arabic-English bilinguals with LEP ($n = 30$; LEP-AR) from Canada, and Canadian NSEs ($n = 24$). The LEP-RO group was established by default: all participants grew up in a non-English-speaking country and learned English later in life. LEP-AR was psychometrically operationalized: a BNT-15 score of ≤11 was required – a level of performance highly specific to LEP status (Ali, Elliott, et al., 2022; Brantuo et al., 2022). The NSE sample included participants born and raised in an English-speaking part of Canada.

To control for noncredible responding as a confound (Abeare et al., 2021), only participants who passed the first trial of the Test of Memory Malingering (i.e., scored >43 on the TOMM-1; Crişan & Erdodi, 2022; Erdodi, 2022; Jones, 2013; Kulas et al., 2014; Rai & Erdodi, 2021) were included in the study. Six participants from LEP-RO and four from LEP-AR were excluded based on their TOMM-1 scores. All NSEs scored above the cutoff and were retained in the study. No participant reported any neurological or neuropsychological condition associated with cognitive impairment. The three samples were similar in age and gender. LEP-RO participants had higher levels of education than NSEs (Table 1).

Materials

All participants were administered a battery of neuropsychological tests in English, including the first three trials of the Stroop test of the Delis–Kaplan Executive Function System (D-KEFS; Delis et al., 2001), the HVLT-R (Benedict et al., 1998) with the newly developed Forced Choice Recognition (FCR; Abeare et al., 2020; Cutler et al., 2021), the Digit Span and Digit-Symbol Coding (CD) subtests of the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; Wechsler, 1997), the TMT (Reitan, 1955), animal (Gladsjo et al., 1999) and Emotion Word Fluency test (EWFT; Abeare et al., 2017).

The EWFT instructs examinees to generate as many emotion words as possible within 1 minute. The initial validation study placed the normative output (raw score) in Canadian university students between 10.6 ($SD = 3.3$) and 11.4 words ($SD = 3.3$; Abeare et al., 2017). Subsequent research reported slightly higher but more variable performance in cognitively healthy students ($M = 13.3$, $SD = 3.3$) and slightly lower scores in clinical patients ($M = 9.9$, $SD = 4.4$; Abeare An et al., 2022).

Age-corrected scaled scores (ACSSs) for the D-KEFS, HVLT-R, Digit Span, and CD were derived from norms published in the Technical Manuals. Demographically adjusted T-scores for TMT and animal fluency were determined using norms published by Heaton et al. (2004). Although norms developed on and for NSEs in the USA cannot be assumed to be the appropriate reference

Table 1. Sample characteristics

	Samples									Sig.	
	LEP-RO[1]		LEP-AR[2]		NSE[3]						
	n = 59		*n* = 30		*n* = 24						
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	F/χ^2	*p*	η_p^2	*post hocs*	*g* [95% CI]
Age	23.8	6.1	22.0	4.8	21.2	4.9	2.22	.113	.039	–	
Education	14.8	2.2	14.0	1.7	13.1	1.3	6.9	.002	.111	1–3	.86 [.36–1.35]
BNT-15	10.6	2.4	9.0	2.1	14.1	1.0	40.2	<.001	.422	1–2	.69 [.24–1.15]
										2–3	2.99 [2.21–3.77]
										1–3	1.67 [1.13–2.21]
% Female	89.8		83.3		83.3		1.03	.596	–	–	–

Note. BNT-15: Boston Naming Test – Short Form (administered in English); LEP-RO: Romanian Limited English Proficiency sample; LEP-AR: Canadian Arabic LEP sample; NSE: Canadian native speakers of English; η_p^2: Partial Eta-Squared (effect size for ANOVAs); *Sig. post hoc*s: Significant *post hoc* contrasts (Games-Howell tests, *p* < .05); *g*: Effect size (Hedge's *g*); 95% CI: 95% Confidence interval.
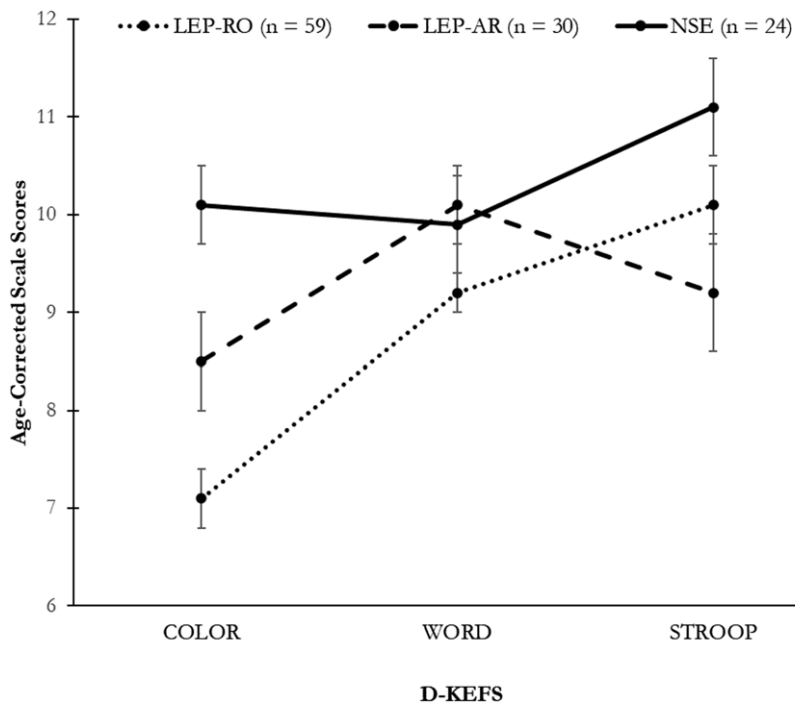

Figure 1. Pattern of performance on various trials of the Delis–Kaplan Executive System (D-KEFS) across the three samples. LEP-RO: Romanian Limited English Proficiency sample (*n* = 59); LEP-AR: Canadian Arabic LEP sample (*n* = 30); NSE: Canadian native speakers of English (*n* = 24). Error bars represent the standard error of the mean.

group for examinees with LEP in the USA, Canada, or other countries, these are the normative data most likely available to clinicians when assessing LEP examinees. Therefore, an empirical evaluation of the extent to which widely used norms may or may not be appropriate for such individuals is directly relevant to North American neuropsychologists.

Procedure

Participants were recruited as volunteers in a study on cognitive performance and received extra credit for their time. Tests were administered face-to-face individually in quiet rooms by bilingual research assistants with a Bachelor's degree in psychology, relevant coursework in psychometrics, and specialized training and ongoing supervision received by the first and last authors in administering and scoring the employed battery. Research assistants in Romania and Canada followed the same standardized procedure developed by test publishers during administration and scoring. All tests were administered in English, following standard protocols. In addition, animal

fluency and EWFT were administered in both languages only in the LEP samples to directly evaluate the effect of language of administration (native versus English). All data collection, storage, and processing were done with the approval of relevant institutional authorities regulating research involving human participants, in compliance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards.

Data analysis

Descriptive statistics (percentage, *M*, *SD*) for each group were reported as relevant. The main inferential statistics evaluating the significance of between-group differences were one-way ANOVAs, chi-square, and independent (Welch's) and within-sample *t*-tests (all contrasts were two-tailed). *Post hoc* contrasts were performed using the Games–Howell test to control the familywise error rate and protect against alpha inflation. Effect size estimates were expressed in Hedge's *g* (with corresponding 95% CIs) and partial eta squared (η_p^2).

Table 2. One-way ANOVAs comparing performance across samples on tests of visuomotor speed, attention, and executive function

Test	Trial	Score	LEP-RO[1] $n = 59$ M	SD	LEP-AR[2] $n = 30$ M	SD	NSE[3] $n = 24$ M	SD	F/t	p	η_p^2	Sig. post hocs	g [95%CI]
Digit span	–	ACSS	7.5	2.6	9.3	2.4	10.2	4.3	8.27	<.001	.131	1–2	.70 [.25–1.16]
												1–3	.84 [.35–1.33]
	LDF	Raw	5.8	1.2	6.3	1.3	6.4	1.6	2.56	.082	.044	–	
	LDB	Raw	4.2	1.0	4.9	1.1	5.0	1.6	5.73	.004	.094	1–2	.67 [.22–1.12]
												1–3	.66 [.17–1.14]
Coding	–	ACSS	10.0	2.7	10.3	2.5	11.6	3.2	3.1	.049	.053	–	
TMT-A	–	T	31.9	9.1	42.8	9.6	47.6	11.0	27.7	<.001	.335	1–2	1.17 [.69–1.64]
												1–3	1.61 [1.07–2.14]
TMT-B	–	T	40.0	10.3	42.2	10.2	48.5	7.8	6.41	.002	.104	1–3	.90 [.41–1.39]
												2–3	.71 [.16–1.26]
TMT B/A	–	Raw	1.85	0.60	2.50	0.74	2.40	0.73	12.3	<.001	.183	1–2	.99 [.53–1.45]
												1–3	.85 [.36–1.34]
D-KEFS	COL	ACSS	7.1	2.4	8.5	2.5	10.1	2.0	14.9	<.001	.214	1–2	.57 [.12–1.02]
												1–3	1.30 [.78–1.81]
												2–3	.69 [.14–1.24]
	WOR	ACSS	9.2	1.8	10.1	2.3	9.9	2.4	2.5	.084	.044	–	
	STR	ACSS	10.1	2.8	9.2	3.3	11.1	2.3	2.9	.059	.050	2–3	.65 [.10–1.2]
Animals	–	Raw	15.4	4.0	17.5	5.3	22.4	5.8	18.2	<.001	.249	1–3	1.51 [.99–2.04]
												2–3	.87 [.31–1.44]
		T	29.8	9.5	34.0	12.8	46.1	11.7	19.0	<.001	.257	1–3	1.59 [1.05–2.12]
												2–3	.97 [.40–1.53]
EWFT	–	Raw	9.3	3.6	8.0	2.6	–	–	3.59	.062	–	–	.39 [.05–.84]
Animals*	–	Raw	21.8	5.0	12.5	5.1	–	–	67.01	<.001	–	–	1.83 [1.33–2.36]
		T	42.7	10.8	21.8	13.9	–	–	51.65	<.001	–	–	1.74 [1.24–2.26]
EWFT*	–	Raw	10.7	3.4	6.2	2.2	–	–	55.96	<.001	–	–	1.46 [.98–1.96]

Note. All tests were administered in English unless marked with * (those tests were administered in the native language of the LEP sample); TMT: Trail Making Test; D-KEFS: Delis–Kaplan Executive Systems; EWFT: Emotion Word Fluency Test; Animals: Category fluency; EWFT: Emotion Word Fluency Test; LDF: Longest digit span forward; LDB: Longest digit span backward; COL: Color Naming; WOR: Word Reading; STR: Stroop; ACSS: Age-corrected scaled score ($M = 10$, $SD = 3$); T: T-score ($M = 50$, $SD = 10$); LEP-RO: Romanian limited English proficiency sample; LEP-AR: Canadian Arabic LEP sample; NSE: Canadian native speakers of English; η_p^2: Partial Eta-Squared (effect size for ANOVAs); Sig. post hocs: Significant post hoc contrasts (Games-Howell tests, $p < .05$); $t =$ Welch's t test; $g =$ Effect size for significant post hoc constrasts (Hedge's g); 95%CI = 95% Confidence interval.

Results

A large main effect on Digit Span ACSS and a medium effect on longest Digit Span backward were driven by the below-average score of LEP-RO. No difference was noted on longest Digit Span forward (Table 2). There was a small-medium main effect on CD caused by the above-average performance of NSE participants. An extremely large effect emerged on TMT-A, driven by the unusually low score of the LEP-RO sample. The performance gap between groups narrowed on the TMT-B but remained significant. A large effect emerged on the TMT B/A raw score ratio, driven by low scores of the LEP-RO sample (indicating better cognitive flexibility relative to visuomotor sequencing speed). A very large main effect was observed on the Color Naming subtest of the D-KEFS, reflecting a linear increase in performance from LEP-RO through LEP-AR to NSE. The contrasts on the Word Reading and Stroop subtests of the D-KEFS were not significant. Figure 1 displays the between-group trends on the three trials of the D-KEFS.

A very large main effect re-emerged on animal fluency in English, driven by the normative performance of the NSE sample relative to the mean score in the impaired and borderline range, respectively, of the LEP samples. The contrast between the two LEP samples on EWFT approached significance (medium effect). When we compared performances of the two LEP samples on animal fluency and EWFT administered in their native language (Romanian and Arabic), extremely large effects emerged for both measures.

Finally, within-sample t-tests revealed a significantly higher performance in raw scores on animal fluency [$t(58) = 12.9$,

$p < .001$, $d = 1.68$, extremely large effect] and EWFT [$t(58) = 3.09$, $p < .01$; $d = 0.40$, medium effect] in Romanian within the LEP-RO sample. At T-score levels, mean performance on animal fluency shifted from the impaired (English) to the low average (Romanian) range [$t(58) = 10.9$, $p < .001$, $d = 1.41$, extremely large effect]. Within the LEP-AR sample, all three contrasts were significant but *in the opposite direction*: participants performed better when animal fluency [$t(29) = -4.97$, $p < .001$] and EWFT [$t(29) = -3.75$, $p < .01$] were administered in English ($d = 0.68–0.91$, large effects). At T-score levels, mean performance on animal fluency shifted from the borderline (English) to the impaired (Arabic) range [$t(29) = -4.90$, $p < .001$; $d = 0.89$, large effect].

Significant main effects emerged on all three individual acquisition trials of the HVLT-R, although the magnitude of the difference declined gradually with each subsequent trial (from large to medium effects). However, a large effect re-emerged on the sum of Trials 1–3 (Table 3). There was a very large effect on delayed free recall. Although the ANOVA remained significant on recognition performance, the effect size was notably smaller (medium) on raw scores. Once age correction was applied (T-scores), between-group differences disappeared. All contrasts above were driven by the notably lower performance of the LEP-AR sample. Although the main effect on the FCR trial was significant, this likely reflects the mathematical artifact of very low SDs, as all three samples performed near the ceiling (i.e., a score of 12.0). Figure 2 provides a visual summary of the between-group patterns of auditory verbal learning performance.

Table 3. One-way ANOVAs comparing performance across samples on the HVLT-R

		Samples										
		LEP-RO[1]		LEP-AR[2]		NSE[3]						
		n = 59		*n* = 30		*n* = 24						
Trial	Score	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	η_p^2	Sig. *post hoc*s	*g* [95% CI]
1	Raw	7.0	1.8	5.6	1.8	7.3	1.3	8.14	.001	.129	1–2	.77 [.32–1.22]
											2–3	1.05 [.48–1.62]
2	Raw	9.1	2.0	8.2	1.7	10.0	1.0	6.83	.002	.110	2–3	1.24 [.65–1.82]
											1–3	.50 [.02–.98]
3	Raw	10.3	1.4	9.7	1.5	10.6	1.2	3.42	.036	.059	2–3	.64 [.09–1.20]
1–3	Raw	26.5	4.3	23.6	4.4	27.9	2.4	8.49	<.001	.134	1–2	.66 [.21–1.11]
											2–3	1.16 [.58–1.74]
	T	44.1	10.9	35.2	10.9	46.1	7.8	9.71	<.001	.150	1–2	.81 [.35–1.26]
											2–3	1.11 [.54–1.69]
DR	Raw	9.6	1.9	7.5	2.6	10.0	1.2	13.7	<.001	.199	1–2	.96 [.50–1.43]
											2–3	1.17 [.59–1.75]
	T	45.1	11.5	33.9	11.7	48.2	8.1	14.2	<.001	.205	1–2	.96 [.50–1.42]
											2–3	1.37 [.78–1.97]
RD	Raw	11.4	0.8	10.6	2.0	11.3	0.8	4.1	.019	.069	–	
	T	53.0	7.0	52.7	7.4	50.0	7.9	1.44	.242	.026	–	
FCR	Raw	12.0	0.1	12.0	.00	11.9	0.3	3.82	.025	.065	–	

Note. All tests were administered in English, following standard instructions; HVLT-R: Hopkins Verbal Learning Test – Revised; 1–3: Acquisition trials (sum of scores across trials 1 through 3); DR: Delayed free recall; RH: Yes/No recognition hits (true positives); RD: Recognition discrimination (true positives minus false positives); FCR: Forced Choice Recognition; LEP-RO: Romanian Limited English Proficiency sample; LEP-AR: Canadian Arabic LEP sample; NSE: Canadian native speakers of English; η_p^2: Partial Eta-Squared (effect size for ANOVAs); *Sig. post hocs*: Significant *post hoc* contrasts (Games-Howell tests, *p* < .05); *g* = Effect size for significant post hoc constrasts (Hedge's *g*); 95%CI = 95% Confidence interval.
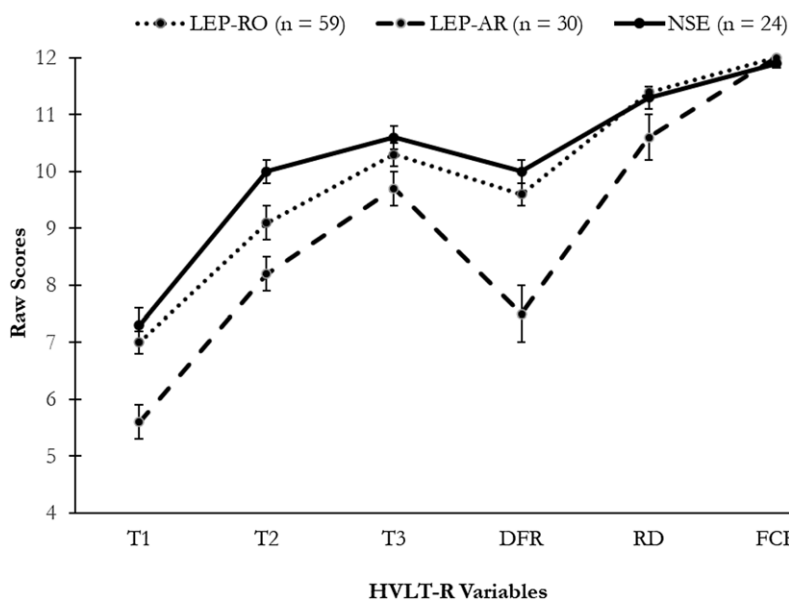

Figure 2. Pattern of performance on various trials of the Hopkins Verbal Learning Test – Revised (HVLT-R) across the three samples. LEP-RO: Romanian Limited English Proficiency sample (*n* = 59); LEP-AR: Canadian Arabic LEP sample (*n* = 30); NSE: Canadian native speakers of English (*n* = 24). T: Trial; DFR: Delayed free recall; RD: Recognition discrimination (true positives minus false positives); FCR: Forced Choice Recognition; Error bars represent the standard error of the mean.

Given the prominence of North American normative systems, one-sample *t*-tests were computed for each of the samples against US norms (Table 4). The LEP-RO performed significantly below the normative mean on Digit Span (large effect), TMT A & B (very large and large effects), animals (very large effect), HVLT-R (medium effects), D-KEFS Color Naming (large effect) and Word Reading (small-medium effect), showing no difference on CD and Stroop. The LEP-AR performed significantly below the normative mean on Digit Span (large effect), TMT A & B (large effects), animals (very large effect), HVLT-R (small to very large effects), and the Color Naming (medium effect) subtest, with no difference on Digit Span, CD, and Word Reading or Stroop subtests. The NSE sample performed *above* the normative mean on CD and Stroop

(medium effects) and *below* the normative mean on the acquisition trials of the HVLT-R (medium effect).

Since a BNT-15 score ≤ 11 has been proposed as a psychometric marker of LEP (Ali, Elliott, et al., 2022; Brantuo et al., 2022; Erdodi et al., 2017a), whereas a score of 12 has been identified as the low end of intact performance among NSEs (Abeare et al., 2022), two subgroups were created first within the LEP-RO sample along this cutoff. Participants with BNT-15 ≤ 11 scored significantly lower than those with BNT-15 ≥ 12 on animal fluency in *both languages* (despite smaller effects during the Romanian administration) and the English administration of the EWFT (large effect). Similarly, large effects emerged on the time-to-completion of both the Yes/No and the FCR recognition trials of the HVLT-R (Table 5).

Table 4. One sample t-tests against the normative mean across samples.

			Sample								
			LEP-RO[1]			LEP-AR[2]			NSE[3]		
			$n = 59$			$n = 30$			$n = 24$		
Test	Trial	Score	t	p	g [95% CI]	t	p	g [95% CI]	t	p	g [95% CI]
Digit Span	–	ACSS	−7.42	<.001	.87 [0.53–1.21]	−1.50	.145	.24 [−.17 to .65]	0.19	.851	.06 [−.51 to .39]
Coding	–	ACSS	−0.05	.961	.00 [−.32 to .32]	0.74	.465	.10 [−.32 to .31]	2.48	.021	.53 [.08–.98]
TMT-A	–	T	−15.4	<.001	1.88 [1.49–2.25]	−4.11	<.001	.73 [.31–1.14]	−1.06	.300	.24 [−.21 to .68]
TMT-B	–	T	−7.44	<.001	.99 [.65–1.33]	−4.21	<.001	.78 [.36–1.2]	−0.94	.355	.16 [−.29 to .60]
Animals		T	−16.3	<.001	2.06 [1.66–2.45]	−6.85	<.001	1.49 [1.05–1.94]	−1.62	.119	.38 [−.07 to .83]
HVLT-R	1–3	T	−4.14	<.001	.57 [.24–.90]	−7.47	<.001	1.45 [1.01–1.89]	−2.44	.023	.41 [−.04 to .85]
	DR	T	−3.32	.002	.46 [.14–.79]	−7.59	<.001	1.55 [1.1–1.99]	−1.10	.281	.19 [−.26 to .63]
	RD	T	3.28	.002	.33 [.01–.66]	1.96	.059	.29 [−.69 to .12]	0.03	.98	.00 [−.45 to .45]
D-KEFS	COL	ACSS	−9.54	<.001	1.03 [.69–1.37]	−3.27	.003	.52 [.11–.93]	0.20	.842	.04 [−.48 to .41]
	WOR	ACSS	−3.62	.001	.31 [.02–.63]	0.31	.756	.03 [−.44 to .37]	−0.26	.801	.04 [−.41 to .48]
	STR	ACSS	0.28	.782	.03 [.29–.36]	−1.26	.219	.26 [−.15 to .67]	2.37	.027	.38 [−.83 to .07]

Note. All tests were administered in English, following standard instructions; TMT: Trail Making Test; Animals: Category fluency; HVLT-R: Hopkins Verbal Learning Test – Revised; D-KEFS: Delis–Kaplan Executive Systems; Animals: Category fluency; 1–3: Acquisition trials (sum of scores across trials 1 through 3); DR: Delayed free recall; RD: Recognition discrimination (true positives minus false positives); COL: Color Naming; WOR: Word Reading; STR: Stroop; ACSS: Age-corrected scaled score ($M = 10$, $SD = 3$); T: T-scores ($M = 50$; $SD = 10$); LEP-RO: Romanian Limited English Proficiency sample; LEP-AR: Canadian Arabic LEP sample; NSE: Canadian native speakers of English; g = Effect size (Hedge's g); 95%CI = 95% Confidence interval.

Within-sample *t*-tests revealed that LEP-RO participants with BNT-15 ≤ 11 performed better at raw score level on both animal fluency [$t(35) = 9.89$, $p < .001$; $d = 1.65$, very large effect] and EWFT [$t(35) = 3.75$, $p < .01$; $d = 0.63$, medium effect] administered in Romanian. Their mean animal fluency T-score shifted from the impaired (English) to the low average (Romanian) range [$t(35) = 8.38, p < .001; d = 1.40$, extremely large effect]. Similar results were observed in participants with BNT-15 ≥ 12 on animal fluency raw [$t(22) = 8.51$, $p < .001$; $d = 1.77$] and T-scores [$t(22) = 7.17$, $p < .001$; $d = 1.50$], with extremely large effects. The mean animal fluency T-score shifted from the borderline (English) to the average (Romanian) range. However, there was no difference in EWFT performance within this subset of the LEP-RO sample as a function of the administration language [$t(22) = .08$, $p = .945$].

To control for the method variance in selecting participants for the LEP-RO (by default) and the LEP-AR (BNT-15 ≤ 11) samples, the main contrasts were re-computed after Romanian participants with BNT-15 scores >11 were excluded. This change in the composition of the LEP-RO sample ensured that the two groups had comparable levels of English proficiency. The overall pattern of positive and negative findings captured in Tables 2 and 3 was preserved after equalizing the groups (Table 6).

Finally, to investigate whether there is an incremental loss in performance on cognitive tests as a function of decreasing English proficiency, test scores were compared across five BNT-15 scores: 11, 10, 9, 8, and ≤7 using a series of one-way ANOVAs (Table 7). Only two significant main effects emerged: on CD ($\eta_p^2 = .205$, large) and animal fluency T-scores ($\eta_p^2 = .149$, large). Examining the pattern of CD scores revealed that the finding was driven by the combination of an isolated high average range mean associated with BNT-15 = 11 (12.1) compared to a narrow (average) range performance ($M = 9.5$–9.8) at the other four levels of BNT-15 and low variability ($SD = 1.9$–2.5). However, a linear decline in animal fluency T-scores was observed, from $M = 35.6$ at BNT-15 = 11 to $M = 24.8$ at BNT-15 ≤ 7.

When developmental history and cognitive profile collide: a case study

Although learning a language outside the sensitive period (age > 15) is commonly considered a developmental marker of LEP (Johnson & Newport, 1989; Lenneberg, 1967; Sakai, 2005), individual variability in language acquisition results in notable exceptions from this principle. To illustrate this, we present psychometric data from a a 47-year-old right-handed female patient with 16 years of education referred to the senior author's private practice for assessment following an uncomplicated mild traumatic brain injury. She grew up speaking Russian, immigrated to Canada at age 18, and obtained a bachelor's degree. By history, she would be classified as LEP. However, she had no obvious accent when speaking English and obtained the following scores on verbal neuropsychological tests: BNT-15 = 14 (the mean of the NSE sample in the present study was 14.1 and 13.5 in the most recently published norms; Abeare et al., 2022); Complex Ideational Material = 12 (perfect score); letter and animal fluency $T = 61$; California Verbal Learning Test acquisition trials raw score = 66/80 ($T = 69$), long-delay free recall raw score = 4/16 (z-score = 1.0); Similarities ACSS = 16, Vocabulary ACSS = 19 (Verbal Comprehension Index = 150). Based on her cognitive profile, her neuropsychological functioning better matches an NSE's.

Discussion

This study was designed to investigate geographic differences in cognitive profiles associated with LEP and compare them to norms developed on and for NSEs. To this end, two different LEP samples were recruited (Romanian and Arabic Canadian students), and their cognitive profiles were compared to NSE norms and a student sample of Canadian NSEs. We predicted that LEP participants would perform worse than NSEs and below the normative mean on verbal tests; no difference between NSE and LEP on nonverbal tests; and that performance on verbal tests would differ based on English proficiency levels within the LEP sample. Results generally supported the first hypothesis, with several notable exceptions: the LEP-RO and LEP-AR samples demonstrated a unique pattern of strengths and weaknesses that defies a unifying interpretation. The support for the second hypothesis was mixed due to the divergent performance between the two LEP samples. The third hypothesis was only supported in the verbal fluency tests and the HVLT-R time-to-completion metrics.

Results are broadly consistent with previous research on the deleterious effect of LEP on performance during verbal tasks

Table 5. Performance on cognitive tests within the Romanian LEP sample as a function of BNT-15 score

Test	Score	LA	Scale	BNT-15 Score ≤11 n = 36		BNT-15 Score >11 n = 23		t	p	g [95% CI]
				M	SD	M	SD			
Digit Span	Total	EN	Raw	14.0	3.9	15.4	2.4	2.91	.094	.41 [.12–.93]
	Total	EN	ACSS	7.0	2.9	8.3	1.9	4.24	.044	.50 [.03–1.03]
	LDF	EN	Raw	5.7	1.3	5.8	1.0	0.12	.735	.08 [.44–.61]
	LDB	EN	Raw	4.1	1.0	4.5	1.0	3.11	.084	.39 [.13–.92]
Coding	–	EN	ACSS	10.2	2.5	9.6	3.0	0.69	.411	.22 [−.31 to .74]
TMT	A	EN	T	31.9	9.2	31.8	9.1	0.04	.947	.01 [−.51 to .53]
	B	EN	T	40.1	10.9	39.9	9.5	0.04	.950	.02 [−.50 to .65]
Animals	–	RO	Raw	20.4	4.5	23.8	5.1	6.91	.012	.71 [.17–1.25]
		RO	T	39.8	10.7	47.1	9.6	7.42	.009	.70 [.16–1.24]
		EN	Raw	13.8	3.0	18.0	4.0	19.1	<.001	1.21 [.64–1.78]
		EN	T	26.0	7.8	35.8	9.0	18.6	<.001	1.17 [.60–1.73]
EWFT	–	RO	Raw	10.4	3.4	11.1	3.4	0.73	.398	.20 [−.73 to .32]
		EN	Raw	8.1	2.6	11.1	4.2	9.1	.005	.89 [.35–1.44]
D-KEFS	COL	EN	ACSS	6.6	2.3	7.7	2.4	3.2	.080	.46 [.07–.99]
	WOR	EN	ACSS	9.0	1.9	9.4	1.6	0.72	.400	.22 [−.30 to .75]
	STR	EN	ACSS	9.8	2.8	10.6	2.8	1.23	.274	.28 [−.24 to .81]
HVLT-R	Trial 1	EN	Raw	6.9	1.8	7.2	1.7	.36	.550	.17 [−.69 to .36]
	Trial 2	EN	Raw	8.9	2.2	9.5	1.6	1.60	.212	.30 [−.82 to .23]
	Trial 3	EN	Raw	10.1	1.6	10.7	0.9	2.5	.120	.43 [−.96 to .10]
	1–3	EN	Raw	25.9	4.7	27.4	3.6	1.82	.182	.34 [−.87 to .18]
		EN	T	43.0	11.5	45.9	9.9	1.07	.305	.26 [−.79 to .26]
	DR	EN	Raw	9.3	2.0	10.1	1.6	2.90	.094	.43 [−.95 to .10]
		EN	T	42.9	12.0	48.4	9.9	3.52	.066	.48 [−1.01 to .05]
	RD	EN	Raw	11.4	0.8	11.3	0.8	0.22	.642	.12 [−.40 to .65]
		EN	T	53.8	6.2	51.7	8.1	1.22	.276	.30 [−.23 to .82]
		EN	T2C	49.1	7.8	42.7	7.8	9.54	.003	.81 [.27–1.35]
	FCR	EN	Raw	12.0	0.2	12.0	0.0	–	–	.00 [−.52 to .52]
		EN	T2C	36.8	6.7	31.8	6.6	7.75	.008	.74 [.20–1.28]

Note. TMT: Trail Making Test; EWFT: Emotion Word Fluency Test; D-KEFS: Delis–Kaplan Executive Function System; HVLT-R: Hopkins Verbal Learning Test – Revised; LDF: Longest digits forward; LDB: Longest digits backward; COL: Color Naming; WOR: Word Reading; STR: Stroop; 1–3: Acquisition trials (sum of scores across trials 1 through 3); DR: Delayed-free recall; RD: Recognition discrimination (true positives minus false positives); FCR: Forced Choice Recognition; LA: Language of administration; EN: English; RO: Romanian; ACSS: Age-corrected scaled score; T: Demographically adjusted T-score based on norms by Heaton et al. (2004); T2C: Time to completion (seconds); *t* = Welch's *t* test; *g* = Effect size (Hedge's *g*); 95% CI = 95% Confidence interval.

(Bialystok et al., 2008, 2009; Boone et al., 2007; Coderre et al., 2013; Kisser et al., 2012; Mattys et al., 2017; Rivera Mindt et al., 2008; Walker et al., 2010). Previous reports of the heightened sensitivity of the D-KEFS Color Naming to LEP relative to Word Reading were replicated (Brantuo et al., 2022), with one caveat: LEP-RO continued to improve on the Stroop task, whereas performances of LEP-AR declined. Consistent with existing research (Brantuo et al., 2022; Erdodi et al., 2017a), animal fluency was very sensitive to LEP, as evidenced by a mean performance of 1.5–2 SDs below the normative mean. Consistent with previous reports (Wauters & Marquardt, 2017), the EWFT was less susceptible to the administration language than animal fluency, although both the magnitude and the direction of the effect of native versus English administration were different in LEP-RO from LEP-AR.

The performance of the LEP-RO sample improved during the native language compared to the English administration of the animal fluency test. Applying the demographically adjusted norms by Heaton et al. (2004) to raw scores increased their average scores by almost 1.5 SDs. However, the LEP-AR sample demonstrated the opposite pattern: participants performed better during the English administration, resulting in a 1 SD difference. This pattern complicates the interpretation of the results and precludes clear recommendations to assessors in clinical settings. Findings from the LEP-RO sample indicate that scores during the task's standard English administration underestimate semantic fluency skills that could be obtained in their native language by 1–1.5 SDs. Therefore,

adjusting the T-score obtained in English by 10–15 T-score points may provide a more accurate estimate of the true cognitive ability of LEP examinees who could not be tested in their native language.

However, findings in the LEP-AR sample suggest that such an adjustment is far from universally applicable. Whether the Heaton norms provide a valid normative comparison for individuals with LEP has yet to be established. Known variability in verbal fluency scores as a function of broader cultural and linguistic variables (Ardila, 2020) suggests that the accurate clinical interpretation of test scores may require a deeper understanding of the complex interactions among the various factors influencing performance on cognitive testing.

Similar to the clinical case study, the LEP-RO sample produced an auditory verbal memory profile that was indistinguishable from that of NSEs, whereas the LEP-AR consistently underperformed the NSE sample. The fact that LEP-AR participants were immersed in an English-speaking language environment, whereas LEP-RO participants lived in a non-English-speaking country, makes this pattern even more difficult to interpret. The most parsimonious explanation seems to be the inclusion criterion of BNT-15 ≤ 11: although needed to ensure that the English-Arabic bilinguals had LEP, it may have inadvertently resulted in oversampling participants from the lower end of the English proficiency continuum.

However, ANOVAs using five levels of the BNT-15 (11, 10, 9, 8, and ≤7) as the independent variable only found two significant

Table 6. Performance on cognitive tests between the Romanian participants with BNT-15 ≤ 11 and the Arabic LEP participants

				Sample						
				RO-LEP		AR-LEP				
				n = 36		n = 30				
Test	Score	LA	Scale	M	SD	M	SD	t	p	g [95% CI]
Digit Span	Total	EN	Raw	14.0	3.9	16.5	3.6	6.89	.011	.66 [.16–1.15]
	Total	EN	ACSS	7.0	2.9	9.3	2.4	12.9	.001	.86 [.35–1.36]
	LDF	EN	Raw	5.7	1.3	6.3	1.3	2.84	.097	.43 [−.06 to .92]
	LDB	EN	Raw	4.1	1.0	4.9	1.1	10.9	.002	.78 [.28–1.29]
Coding	–	EN	ACSS	10.2	2.5	10.3	2.5	0.03	.856	.04 [−.45 to .52]
TMT	A	EN	T	31.9	9.2	42.8	9.6	22.0	<.001	1.15 [.63–1.67]
	B	EN	T	40.1	10.9	42.2	10.2	0.66	.418	.20 [−.29 to .68]
Animals	–	NA	Raw	20.4	4.5	12.5	5.1	44.1	<.001	1.63 [1.07–2.19]
		NA	T	39.8	10.7	21.8	13.9	33.7	<.001	1.45 [.91–2.00]
		EN	Raw	13.8	3.0	17.5	5.2	11.6	.001	.88 [.37–1.39]
		EN	T	26.0	7.8	34.0	12.8	9.1	.004	.76 [.26–1.26]
EWFT	–	NA	Raw	10.4	3.4	6.2	2.2	36.3	<.001	1.42 [.88–1.96]
		EN	Raw	8.1	2.6	8.0	2.6	0.03	.868	.04 [−.45 to .52]
D-KEFS	COL	EN	ACSS	6.6	2.3	8.5	2.5	10.1	.002	.78 [.28–1.29]
	WOR	EN	ACSS	9.0	1.9	10.1	2.3	4.52	.038	.52 [.03–1.01]
	STR	EN	ACSS	9.8	2.8	9.2	3.3	0.50	.481	.20 [−.29 to .68]
HVLT-R	Trial 1	EN	Raw	6.9	1.8	5.6	1.8	7.94	.006	.71 [.21–1.21]
	Trial 2	EN	Raw	8.9	2.2	8.2	1.7	1.88	.175	.35 [−.14 to .84]
	Trial 3	EN	Raw	10.1	1.6	9.7	1.5	1.32	.254	.25 [−.23 to .74]
	1–3	EN	Raw	25.9	4.7	23.6	4.4	4.28	.043	.50 [.01–.99]
		EN	T	43.0	11.5	35.2	10.8	8.04	.006	.69 [.19–1.19]
	DR	EN	Raw	9.3	2.0	7.5	2.6	9.23	.004	.78 [.27–1.28]
		EN	T	42.9	12.0	33.9	11.6	9.65	.003	.75 [.25–1.25]
	RD	EN	Raw	11.4	0.8	10.6	2.0	4.23	.047	.53 [.04–1.02]
		EN	T	53.8	6.2	52.7	7.4	0.47	.496	.16 [−.32 to .65]
	FCR	EN	Raw	12.0	0.2	12.0	0.0	–	–	.00 [−.52 to .52]

Note. TMT: Trail Making Test; EWFT: Emotion Word Fluency Test; D-KEFS: Delis–Kaplan Executive Function System; HVLT-R: Hopkins Verbal Learning Test – Revised; LDF: Longest digits forward; LDB: Longest digits backward; COL: Color Naming; WOR: Word Reading; STR: Stroop; 1–3: Acquisition trials (sum of scores across trials 1 through 3); DR: Delayed free recall; RD: Recognition discrimination (true positives minus false positives); FCR: Forced Choice Recognition; LA: Language of administration; EN: English; NA: Native language (Romanian for the RO and Arabic for the AR participants); ACSS: Age-corrected scaled score; T: Demographically adjusted T-score based on norms by Heaton et al. (2004); T2C: Time to completion (seconds); t = Welch's t test; g = Effect size (Hedge's g); 95% CI = 95% Confidence interval.

contrasts, indicating that below the LEP cutoff (≤11) BNT-15 scores no longer predict performance on most cognitive tests. Therefore, the unexpectedly high performance of the LEP-RO sample cannot be attributed to 23 of the Romanian participants having scored above this cutoff and, hence, proved superior English proficiency than LEP-AR.

Findings on non-verbal tests are less conclusive: although both LEP samples performed close to the normative mean on CD, consistent with previous research (Walker et al., 2010), NSEs scored above it, suggesting that a mild LEP disadvantage persists even in the absence of frank deficits. The outcome on the TMT is puzzling and contradicts previous reports (Boone et al., 2007; Kisser et al., 2012). The LEP-RO sample performed 2 SDs below the normative mean on TMT-A and one SD below on TMT-B. In the context of intact performance on CD and D-KEFS Stroop, these findings are difficult to interpret and serve as an important reminder of the relevance of population-specific norms (Bezdicek et al., 2012, 2016).

Assuming normative performance in examinees with LEP on nonverbal tests on rational grounds alone increases the risk of significant errors in the clinical interpretation of scores (Celik et al., 2020; Funes et al., 2016; Gasquoine & Gonzales, 2012). In fact, our results challenge the notion of "LEP profile" as a unitary construct. They suggest that other parameters (geographic location, level of English proficiency, native language, cultural differences in the significance of response speed, etc.) may be equally important

factors in understanding the clinical implications of test scores by LEP examinees (Ardila, 2020; Coderre et al., 2013; Durand-Lopez, 2020; Marian et al., 2013; Roselli et al., 2002; Singh & Mishra, 2013; Tse & Altarriba, 2012; Walker et al., 2010).

Separating the LEP-RO sample into *high* and *low* English proficiency levels operationalized using BNT-15 scores (Ali, Elliott, et al., 2022; Brantuo et al., 2022) revealed a performance pattern with potential clinical relevance. Although both groups obtained significantly lower scores during the English relative to Romanian administration of animal fluency, participants with BNT-15 ≥ 12 performed consistently better on *both* administrations. These findings support the use of the BNT-15 as an objective index of English proficiency (Erdodi et al., 2017a) and reveal that BNT-15 scores may tap the broader construct of general verbal skills independent of any specific language, which includes fund of word knowledge and the speed of lexical retrieval. In other words, BNT-15 preserves its original function of measuring cognitive functioning in addition to LEP status.

Finally, a BNT-15 ≤ 11 score was associated with higher time-to-completion on the HVLT-R recognition trials, indicating increased processing demands in participants with lower levels of English proficiency. This finding has implications for both performance validity assessment and academic accommodations for LEP students at English-speaking institutions. Since time-to-completion often serves as an index of response credibility on word recognition tests generally (Cutler et al., 2022; Erdodi &

Table 7. Test performance across five levels of BNT-15 among Romanian and Arabic participants with BNT-15 scores ≤ 11 (n = 66)

		Test																				
		Digit Span		Coding		TMT-A		TMT-B		Animals		EWFT		HVLT 1-3		HVLT DR		COL_{D-KEFS}		WOR_{D-KEFS}		STR_{D-KEFS}	
		ACSS		ACSS		T-score		T-score		T-score		Raw		T-score		T-score		ACSS		ACSS		ACSS	
BNT-15	n	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
11	18	8.2	2.7	12.1	2.3	38.2	9.4	42.9	12.5	35.6	11.0	9.0	2.8	44.1	11.2	43.8	10.5	8.2	2.3	10.6	2.1	10.3	2.8
10	15	8.9	3.3	9.7	1.9	39.0	11.3	37.9	11.7	31.0	12.9	7.9	1.8	39.9	12.0	37.5	14.2	7.2	2.9	9.3	2.5	9.3	3.9
9	10	6.2	2.5	9.5	2.5	32.8	7.4	41.7	12.0	26.0	6.5	8.1	3.0	39.6	9.4	41.9	14.5	7.2	2.4	9.2	1.5	9.6	2.3
8	9	8.9	3.1	9.8	2.4	39.3	12.8	43.8	8.3	27.1	11.5	7.9	2.6	37.1	16.6	34.6	13.9	7.2	3.8	9.6	2.7	8.6	3.2
≤7	14	7.6	2.7	9.5	2.2	34.4	12.5	39.8	6.4	24.8	8.3	7.2	2.5	34.5	9.0	34.4	9.7	7.1	1.6	9.2	2.2	9.4	2.8
F		1.68		3.93		.87		.68		2.67		1.02		1.44		1.64		.53		.41		.57	
p		.167		.007		.488		.607		.041		.405		.231		.176		.718		.802		.685	
η_p^2		.099		.205		.054		.043		.149		.063		.086		.097		.033		.026		.036	

Note. All tests were administered in English; BNT-15: Boston Naming Test – Short Form; TMT: Trail Making Test; HVLT: Hopkins Verbal Learning Test – Revised; 1–3: Acquisition trials (sum of performance across Trials 1 through 3); DR: Delayed Recall; COL: Color Naming; WOR: Word Reading; STR: Stroop task; D-KEFS: Delis-Kaplan Executive Function Systems; ACSS: Age-corrected scaled score.

Lichtenstein, 2021; Erdodi et al., 2017b; Kim et al., 2010; Lupu et al., 2018) and the HVLT-R specifically (Cutler et al., 2021), assessors should exercise caution before interpreting slow responding on the HVLT-R as evidence of invalid performance in LEP examinees to protect them against increased false positives. In an academic context, extending the time limit on exams may be construed as a reasonable and necessary accommodation for LEP students (Ali, Brantuo, et al., 2022).

It is widely accepted that translating and norming commonly used neuropsychological tests to all languages is not feasible (Franzen et al., 2021). Administering tests in the examinee's native language is often considered the next best solution for neutralizing the effects of LEP (Franzen et al., 2021; Fujii, 2018). However, our results indicate that such an accommodation can have the opposite (i.e., suppressing rather than enhancing) effect. Indeed, while the Romanian administration significantly improved verbal fluency performance in LEP-RO compared to the English administration, the Arabic administration of these tests produced lower scores in LEP-AR compared to the English administration. This finding suggests that administering psychometric tests in the examinee's native language fails to neutralize LEP as a confound and may even inadvertently magnify distortions within the neurocognitive profile, especially in the absence of appropriate norms for many LEP populations.

Results point towards identifying a list of tests that are robust to the variability in the level of English proficiency as the best pragmatic safeguard to LEP status. Within the present study, three such tests emerged as possible "LEP-resistant" candidates: CD, the Word Reading subtest of the D-KEFS, and the EWFT. Age-corrected T-scores for the Yes/No Recognition Discrimination trial of the HVLT-R were also immune to LEP. However, their utility as an overall measure of auditory verbal learning and memory might be limited, considering that the test's key trials remain vulnerable to LEP.

Results should be interpreted in the context of the study's limitations. The most obvious one is the relatively small samples of convenience. In addition, all participants were recruited from two universities, raising questions about the representativeness of the samples. On the one hand, university students may be cognitively higher functioning than the general population. As such, results may not generalize to clinical populations (Braw, 2021). On the other hand, the significant variability in English proficiency within LEP-RO may have masked general trends relevant to cross-cultural neuropsychology. Additionally, several poorly understood cultural and educational differences between samples might have confounded results, especially on verbal fluency tests (Ardila, 2020). In the absence of appropriate norms for individuals with LEP in general (let alone specific cultural/linguistic communities), the clinical interpretation of cognitive profiles in such populations remains uncertain.

The study also has several strengths. It recruited two LEP samples from different countries (indeed, continents) with linguistically and orthographically dissimilar native languages to empirically investigate the variability in cognitive profiles across different LEP subtypes. Such a design enabled several population- and instrument-specific discoveries with potential clinical and cross-cultural relevance. Participants were screened for non-credible responding, a significant source of error variance in academic research on university students (An et al., 2017; Hurtubise et al., 2020; Roye et al., 2019) and even in normative samples (Erdodi & Lichtenstein, 2017). The battery was selected to include a strategic combination of tests with low and high verbal

mediation informed by previous research to further flesh out LEP-specific performance patterns.

Conclusions

Results are broadly consistent with previous research on the deleterious effects of LEP on cognitive profiles – especially on verbal tests. At the same time, findings revealed clinically significant heterogeneity among individuals with LEP, both within and across samples. Therefore, results challenge the notion that LEP status is a unitary construct and emphasize the importance of population-specific research, as findings may not generalize to different groups with LEP (Braw, 2021). Although the BNT-15 proved a valid overall psychometric marker of English proficiency, some of the evidence suggests that it may also capture general verbal/cognitive skills that are not English-specific. Even in the context of high accuracy scores, LEP is associated with slowed processing speed with clear implications for performance validity assessment and eligibility for academic accommodations. Finally, there may be no straightforward definition of LEP status, as individual history of language acquisition and performance-based markers of English proficiency can produce contradictory conclusions (as illustrated by the case study). More research is needed to better understand cognitive profiles associated with LEP and the optimal method for operationalizing the construct itself.

Conflicts of interest. The authors have no conflicts of interest to declare.

References

Abeare, C., Sabelli, A., Taylor, B., Holcomb, M., Dumitrescu, C., Kirsch, N., & Erdodi, L. (2019). The importance of demographically adjusted cutoffs: Age and education bias in raw score cutoffs within the Trail Making Test. *Psychological Injury and Law*, 12, 170–182. https://doi.org/10.1007/s12207-019-09353

Abeare, C. A., An, K., Tyson, B., Holcomb, M., Cutler, L., May, N., & Erdodi, L. A. (2022). The emotion word fluency test as an embedded performance validity indicator – Alone and in a multivariate validity composite. *Applied Neuropsychology: Child*, 11(4), 713–724. https://doi.org/10.1080/21622965.2021.1939027

Abeare, C. A., Freund, S., Kaploun, K., McAuley, T., & Dumitrescu, C. (2017). The Emotion Word Fluency Test (EWFT): Initial psychometric, validation, and physiological evidence in young adults. *Journal of Clinical and Experimental Neuropsychology*, 39(8), 738–752. https://doi.org/10.1080/13803395.2016.1259396

Abeare, C. A., Hurtubise, J. L., Cutler, L., Sirianni, C., Brantuo, M., Makhzoum, N., & Erdodi, L. A. (2020). Introducing a forced choice recognition trial to the Hopkins Verbal Learning Test – Revised. The Clinical Neuropsychologist. Advance online publication. https://doi.org/10.1080/13854046.2020.1779348

Abeare, K., Cutler, L., An, K. Y., Razvi, P., Holcomb, M., & Erdodi, L. A. (2022). BNT-15: Revised performance validity cutoffs and proposed clinical classification ranges. *Cognitive and Behavioral Neurology*, 35, 155–168. https://doi.org/10.1097/WNN.0000000000000304

Abeare, K., Romero, K., Cutler, L., Sirianni, C. D., & Erdodi, L. A. (2021). Flipping the script: Measuring both performance validity and cognitive ability with the forced choice recognition trial of the RCFT. *Perceptual and Motor Skills*, 128(4), 1373–1408. https://doi.org/10.1177/00315125211019704

Ali, S., Brantuo, M. A., Cutler, L., Kennedy, A., & Erdodi, L. (2022). Limited English proficiency inhibits auditory verbal learning in cognitively intact young adults – Exploring culturally responsive diagnostic and educational safeguards. *Applied Neuropsychology: Child*, 12(2), 97–103. https://doi.org/10.1080/21622965.2022.2034628

Ali, S., Crişan, I., Abeare, C., & Erdodi, L. (2022). Cross-cultural performance validity testing: Managing false positives in Examinees with Limited English proficiency. *Developmental Neuropsychology*, 47(6), 273–294. https://doi.org/10.1080/87565641.2022.2105847

Ali, S., Elliott, L., Biss, R., Abumeeiz, M., Brantuo, M., Kuzmenka, P., Odenigbo, P., & Erdodi, L. (2022). The BNT-15 provides an accurate measure of English proficiency in cognitively intact bilinguals – A study in cross-cultural assessment. *Applied Neuropsychology: Adult*, 29(3), 351–363. https://doi.org/10.1080/23279095.2020.1760277

An, K. Y., Kaploun, K., Erdodi, L. A., & Abeare, C. A. (2017). Performance validity in undergraduate research participants: A comparison of failure rates across tests and cutoffs. *The Clinical Neuropsychologist*, 31(1), 193–206. https://doi.org/10.1080/13854046.2016.1217046

Antoniou, M. (2019). The advantages of bilingualism debate. *Annual Review of Linguistics*, 5, 1–21. https://doi.org/10.1146/annurev-linguistics-011718-011820

Ardila, A. (2020). A cross-linguistic comparison of category verbal fluency test (ANIMALS): A systematic review. *Archives of Clinical Neuropsychology*, 35(2), 213–225. https://doi.org/10.1093/arclin/acz060

Benedict, R. H. B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins Verbal Learning Test – revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12(3), 43–55.

Bezdicek, O., Motak, L., Axelrod, B. N., Preiss, M., Nikolai, T., Vyhnalek, M., Poreh, A., & Ruzika, E. (2012). Czech version of the Trail Making Test: Normative data and clinical utility. *Archives of Clinical Neuropsychology*, 27(8), 906–914. https://doi.org/10.1093/arclin/acs084

Bezdicek, O., Motak, L., Schretlen, D. J., Preiss, M., Axelrod, B. N., Nikolai, T., Peña, J. , Ojeda, N., & Ruzika, E. (2016). Sociocultural and language differences on the Trail Making Test. *Archives of Assessment Psychology*, 6, 33–48.

Bialystok, E., Craik, F. I. M., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest*, 10(3), 89–129.

Bialystok, E., Craik, F. I. M., & Luk, G. (2008). Lexical access in bilinguals: Effects of vocabulary size and executive control. *Journal of Neurolinguistics*, 21, 522–538.

Boone, K., Victor, T. L., Wen, J., Razani, J., & Ponton, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*, 22, 355–365. https://doi.org/10.1016/j.acn.2007.01.010

Brantuo, M., An, K., Biss, R., Ali, S., & Erdodi, L. A. (2022). Neurocognitive profiles associated with limited English proficiency in cognitively intact adults. *Archives of Clinical Neuropsychology*, 37(7), 1579–1600. https://doi.org/10.1093/arclin/acac019

Braw, Y. (2021). Cultural aspects in assessing malingering detection. In A. M. Horton, Jr. & C. R. Reynolds (Eds.), *Detection of Malingering During Head Injury Litigation* (pp. 177–200). Springer.

Celik, S., Kokje, E., Meyer, P., Frolich, L., & Teichmann, B. (2020). Does bilingualism influence neuropsychological test performance in older adults? A systematic review. *Applied Neuropsychology: Adult*, 29(4), 855–873. https://doi.org/10.1080/23279095.2020.1788032

Coderre, E. L., Van Heuven, W. J. B., & Conklin, K. (2013). The timing and magnitude of Stroop interference and facilitation in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, 16(2), 420–441. https://doi.org/10.1017/S

Crişan, I., & Erdodi, L. (2022). Examining the cross-cultural validity of the test of memory malingering and the Rey 15-item test. *Applied Neuropsychology: Adult*, 1–11. https://doi.org/doi:10.1080/23279095.2022.2064753

Cutler, L., Abeare, C., Messa, I., Holcomb, M., & Erdodi, L. A. (2021). This will only take a minute: Time cutoffs are superior to accuracy cutoffs on the Forced Choice Recognition trial of the Hopkins Verbal Learning Test – revised. *Applied Neuropsychology: Adult*, 29(6), 1425–1439. https://doi.org/10.1080/23279095.2021.1884555

Cutler, L., Greenacre, M., Abeare, C. A., Sirianni, C. D., Roth, R. M., & Erdodi, L. (2022). Multivariate models provide an effective psychometric solution to the variability in classification accuracy of D-KEFS Stroop performance validity cutoffs. *The Clinical Neuropsychologist*, 37(3), 617–649. https://doi.org/10.1080/13854046.2022.2073914

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System: Examiner's Manual*. The Psychological Corporation.

Durand-Lopez, E. M. (2020). A bilingual advantage in memory capacity: Assessing the roles of proficiency, number of languages acquired and age of acquisition. *International Journal of Bilingualism*, *25*, 1–16. https://doi.org/doi:0.1177/1367006920965714

Erdodi, L. (2022). Multivariate models of performance validity: The Erdodi Index captures the dual nature of noncredible responding (continuous and categorical). *Assessment*, 1–19. https://doi.org/10.1177/10731911221101910

Erdodi, L. A., Jongsma, K. A., & Issa, M. (2016). The 15-item version of the Boston Naming test as an index of English proficiency. *The Clinical Neuropsychologist*, *31*(1), 168–178. https://doi.org/10.1080/13854046.2016.1224392

Erdodi, L. A., & Lichtenstein, J. D. (2017). Invalid before impaired: An emerging paradox of embedded validity indicators. *The Clinical Neuropsychologist*, *31*(6-7), 1029–1046. https://doi.org/10.1080/13854046.2017.1323119

Erdodi, L. A., & Lichtenstein, J. D. (2021). Information processing speed tests as PVTs. In K. B. Boone (Ed.), *Assessment of Feigned Cognitive Impairment. A Neuropsychological Perspective* (pp. 218–247). Guilford.

Erdodi, L. A., Nussbaum, S., Sagar, S., Abeare, C. A., & Schwartz, E. S. (2017a). Limited English proficiency increases failure rates on performance validity tests with high verbal mediation. *Psychological Injury and Law*, *10*, 96–103.

Erdodi, L. A., Tyson, B. T., Shahein, A., Lichtenstein, J. D., Abeare, C. A., Pelletier, C. L., Zuccato, B. G., Kucharski, B., & Roth, R. M. (2017b). The power of timing: Adding a time-to-completion cutoff to the Word Choice Test and Recognition Memory Test improves classification accuracy. *Journal of Clinical and Experimental Neuropsychology*, *39*, 369–383. https://doi.org/10.1080/13803395.2016.1230181

Eurostat (2018, September). *65% Know At Least One Foreign Language in the EU*. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180926-1

Franzen, S. & European Consortium on Cross-Cultural Neuropsychology (ECCroN) (2021). Cross-cultural neuropsychological assessment in Europe: Position statement of the European Consortium on Cross-Cultural Neuropsychology (ECCroN). *The Clinical Neuropsychologist*, *36*(3), 546–557. https://doi.org/10.1080/13854046.2021.1981456

Fujii, D. E. (2018). Developing a cultural context for conducting a neuropsychological evaluation with a culturally diverse client: The ECLECTIC framework. *The Clinical Neuropsychologist*, *32*(8), 1356–1392. https://doi.org/10.1080/13854046.2018.1435826

Funes, C. M., Hernandez Rodriguez, J., & Lopez, S. R. (2016). Norm comparisons of the Spanish-language and English-language WAIS-III: Implications for clinical assessment and test adaptation. *Psychological Assessment*, *28*(12), 1709–1715. https://doi.org/10.1037/pas0000302

Gasquoine, P. G. (1999). Variables moderating cultural and ethnic differences in neuropsychological assessment: The case of Hispanic Americans. *The Clinical Neuropsychologist*, *13*(3), 376–383.

Gasquoine, P. G., Croyle, K. L., Cavazos-Gonzalez, C., & Sandoval, O. (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Archives of Clinical Neuropsychology*, *22*(8), 991–1001.

Gasquoine, P. G., & Gonzales, D. (2012). Using monolingual neuropsychological test norms with bilingual hispanic Americans: Application of an individual comparison standard. *Archives of Clinical Neuropsychology*, *27*(3), 268–276.

Gladsjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., & Heaton, R. K. (1999). Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment*, *6*(2), 147–178.

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographically Adjusted Neuropsychological Norms for African American and Caucasian Adults – Professional Manual*. Psychological Assessment Resources (PAR Inc.).

Heaton, R. K., Ryan, L., & Grant, I. (2009). Demographic influences and use of demographically corrected norms in neuropsychological assessment. *Neuropsychological Assessment of Neuropsychiatric and Neuromedical Disorders*, *3*, 127–155.

Hurtubise, J., Baher, T., Messa, I., Cutler, L., Shahein, A., Hastings, M., Carignan-Querqui, M., & Erdodi, L. (2020). Verbal fluency and digit span variables as performance validity indicators in experimentally induced malingering and real world patients with TBI. *Applied Neuropsychology: Child*, *9*(4), 337–354. https://doi.org/10.1080/21622965.2020.1719409

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning. The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99.

Jones, A. (2013). Test of Memory Malingering: Cutoff scores for psychometrically defined malingering groups in a military sample. *The Clinical Neuropsychologist*, *27*, 1043–1059. https://doi.org/10.1080/13854046.2013.804949

Kim, M. S., Boone, K. B., Victor, T., Marion, S. D., Amano, S., Cottingham, M. E., Ziegler, E. A., & Zeller, M. A. (2010). The Warrington Recognition Memory Test for words as a measure of response bias: Total score and response time cutoffs developed on, real world, credible and noncredible subjects. *Archives of Clinical Neuropsychology*, *25*(1), 60–70.

Kisser, J. E., Wendell, C. R., Spencer, R. J., & Waldstein, S. R. (2012). Neuropsychological performance of native versus non-native English speakers. *Archives of Clinical Neuropsychology*, *27*(7), 749–755. https://doi.org/10.1093/arclin/acs082

Kousaie, S., Sheppard, C., Lemieux, M., Monetta, L., & Taler, V. (2014). Executive function and bilingualism in young and older adults. *Frontiers in Behavioral Neuroscience*, *8*, 250. https://doi.org/10.3389/fnbeh.2014.00250

Kulas, J. F., Axelrod, B. N., & Rinaldi, A. R. (2014). Cross-validation of supplemental Test of Memory Malingering Scores as performance validity measures. *Psychological Injury and Law*, *7*, 236–244. https://doi.org/10.1007/s12207-014-9200-4

Lenneberg, E. H. (1967). *Biological Foundations of Language*. Wiley.

Lupu, T., Elbaum, T., Wagner, M., & Braw, Y. (2018). Enhanced detection of feigned cognitive impairment using per item response time measurements in the Word Memory Test. *Applied Neuropsychology: Adult*, *25*(6), 532–542. https://doi.org/10.1080/23279095.2017.1341410

Marian, V., Blumenfeld, H. K., Mizrahi, E., Kania, U., & Cordes, A. K. (2013). Multilingual Stroop performance: Effects of trilingualism and proficiency on inhibitory control. *International Journal of Multilingualism*, *10*(1), 82–104. https://doi.org/10.1080/14790718.2012.708037

Mattys, S. L., Baddeley, A., & Trenkle, D. (2017). Is the superior verbal memory span of Mandarin speakers due to faster rehearsal? *Memory and Cognition*, *46*(3), 361–369. https://doi.org/10.3758/s13421-017-0770-8

Papageorgiou, A., Bright, P., Periche Tomas, E., & Filippi, R. (2019). Evidence against a cognitive advantage in the older bilingual population. *Quarterly Journal of Experimental Psychology*, *72*, 1354–1363. https://doi.org/10.1177/1747021818796475

Rai, J., & Erdodi, L. (2021). The impact of criterion measures on the classification accuracy of TOMM-1. *Applied Neuropsychology: Adult*, *28*(2), 185–196. https://doi.org/10.1080/23279095.2019.161.1613994

Reitan, R. M. (1955). The relation of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology*, *19*(5), 393–394.

Rivera Mindt, M., Arentoft, A., Kubo Germano, K., D'Aquila, E., Scheiner, D., Pizzirusso, M., Sandoval, T. C., & Gollan, T. H. (2008). Neuropsychological, cognitive, and theoretical considerations for evaluation of bilingual individuals. *Neuropsychology Review*, *18*(3), 255–268. https://doi.org/10.1007/s11065-008-9069-7

Roselli, M., Ardila, A., Santisi, M. N., Del Rosario Arecco, M., Salvatierra, J., Conde, A., & Lenis, B. (2002). Stroop effect in Spanish-English bilinguals. *Journal of the International Neuropsychological Society*, *8*, 819–827. https://doi.org/doi:10.1017.S1355617702860106

Roye, S., Calamia, M., Bernstein, J. P., De Vito, A. N., & Hill, B. D. (2019). A multi-study examination of performance validity in undergraduate research participants. *The Clinical Neuropsychologist*, *33*(6), 1138–1155. https://doi.org/10.1080/13854046.2018.1520303

Ryan, C. (2013). Language use in the United States: 2011. In *American Community Survey Reports, ACS-22*. U.S. Census Bureau.

Sakai, K. L. (2005). Language acquisition and brain development. *Science*, *310*(5749), 815–819.

Singh, N., & Mishra, R. K. (2013). Second language proficiency modulates conflict-monitoring in an oculomotor Stroop task: Evidence from Hindi-

English bilinguals. *Frontiers in Psychology*, *4*, 1–10. https://doi.org/10.3389/fpsyg.2013.00322

Tse, C. S., & Altarriba, J. (2012). The effects of first- and second-language proficiency on conflict resolution and goal maintenance in bilinguals: Evidence from reaction time distributional analyses in a Stroop task. *Bilingualism: Language and Cognition*, *15*, 663–676. https://doi.org/10.1017/S1366728912000077

Walker, A., Batchelor, J., Shores, A. E., & Jones, M. (2010). Effects of cultural background on WAIS-III and WMS-III performances after moderate-severe traumatic brain injury. *Australian Psychologist*, *45*, 112–122. https://doi.org/10.1080/00050060903428210

Wauters, L., & Marquardt, T. P. (2017). Category, letter, and emotional verbal fluency in Spanish-English bilingual speakers: A preliminary report. *Archives of Clinical Neuropsychology*, *33*(4), 444–457. https://doi.org/10.1093/arclin/acx063

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). The Psychological Corporation.

Yoo, J., & Kaushanskaya, M. (2012). Phonological memory in bilinguals and monolinguals. *Memory and Cognition*, *40*(8), 1314–1330. https://doi.org/10.3758/s13421-012-0237-x