Disruption of the serial position effect as an early marker of Alzheimer’s disease in Spanish–English bilinguals

Reina Mizrahi1, Oona Cromheecke2, David P. Salmon3 and Tamar H. Gollan4

1Department of Cognitive Science, University of California, San Diego, CA, USA, 2Department of Rehabilitation Sciences, Ghent University, Ghent, Belgium, 3Department of Neurosciences, University of California, San Diego, CA, USA and 4Department of Psychiatry, University of California, San Diego, CA, USA

Abstract

Objectives: The present study examined if disruption of serial position effects in list recall could serve as an early marker of Alzheimer’s disease (AD) in Spanish–English bilinguals. Methods: We tested 20 participants initially diagnosed as cognitively normal or with mild cognitive impairment who declined and eventually received a diagnosis of AD (decliners), and 37 who remained cognitively stable (controls) over at least 2 years. Participants were tested on the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) Word List Learning Test in English or Spanish as part of an annual neuropsychological evaluation. Results: Compared to controls, decliners exhibited significantly reduced recall including reduced primacy scores (i.e., items recalled from the first three list items on Trial 1), whereas recency scores (i.e., items recalled from the last 3 list items on Trial 1) were equivalent in decliners and controls. Further analyses suggested that the sensitivity of the primacy effect to preclinical AD was initially stronger in participants tested in Spanish, a surprising finding given that the CERAD was developed for English speakers. However, in the subsequent year of testing, primacy scores declined to the same level regardless of language of testing. Conclusions: Several list learning measures may facilitate early diagnosis of AD in Spanish–English bilinguals, possibly including the relatively understudied primacy effect. Additional studies are needed to investigate the possibility that linguistic or demographic variables might modulate sensitivity of list learning tests to preclinical AD, which could lead to broader improvements in their utility for early diagnosis of AD in all populations.

Keywords: Alzheimer’s disease; Bilingualism; Serial position effects; Primacy effects; List memory; Neuropsychological tests

(Received 10 November 2022; final revision 2 April 2023; accepted 26 April 2023)

Introduction

Approximately 6.5 million people in the U.S. age 65 and older have Alzheimer’s disease (AD) (Alzheimer’s Association, 2022). This number is expected to more than double by 2060 with a large portion of this growth projected to occur in Hispanics. There is general consensus that neural changes of AD begin prior to the observation of significant clinical symptoms (Sperling et al., 2011) and that subtle cognitive changes occur before a clinical diagnosis of dementia and probable AD can be made with any certainty. The ability to identify these early cognitive changes in a growing and diverse population is imperative as new treatment options are being developed that are likely to be most effective early in the course of AD (Cummings et al., 1994). Thus, the field has shifted to identifying neuropsychological markers of “preclinical” AD to provide earlier diagnosis to inform treatment options. This effort has identified a number of neuropsychological and neuroimaging markers of AD in the preclinical period of the disease (Twamley et al., 2006). It should be noted, however, that much of this research was performed only with majority English-speaking populations, so relatively little is known about the efficacy of cognitive markings of preclinical AD in individuals from diverse cultural backgrounds.

Cultural and linguistic factors can lead individuals to under- or over-perform on various neuropsychological tests increasing the possibility of misdiagnosis (Gasquoine & Gonzalez, 2012; Weissberger et al., 2013). Efforts have been made to identify these biases and to reduce or eliminate them by modifying cognitive tests to make them more appropriate for the target population or by employing culturally and linguistically appropriate control comparison groups (Ardila et al., 1994; Judd et al., 2009; Pedraza & Mungas, 2008; Siedlecki et al., 2010; Weissberger et al., 2013; Weissberger et al., 2017). Weissberger et al. (2013) compared neuropsychological test scores of clinically normal Hispanic and non-Hispanic elderly individuals who had cognitive decline in subsequent years (i.e., decliners or preclinical AD) to respective demographically appropriate cognitively healthy Hispanic or non-Hispanic control groups composed of individuals who remained stable in subsequent years. They found that cognitively stable Hispanic controls obtained lower scores on tests of language, executive function, and some measures of global cognition than non-Hispanic controls. In addition, several tests (e.g., picture naming, Trail-Making Test parts A and B) were performed worse by decliners than controls in both the Hispanic and non-Hispanic groups (marking preclinical AD), but some
language-based tests (i.e., vocabulary, semantic fluency) were sensitive to preclinical AD only in non-Hispanic groups.

Performance on language-based tests might be especially likely to vary by cultural group given the difficulty of producing translation equivalent versions of tests and the effects of bilingualism on test performance (Peña, 2007; for review, see Kroll & Gollan, 2013). On confrontation naming tests, for example, bilinguals name pictures more slowly than monolinguals, have more tip-of-the-tongue states (Gollan & Silverberg, 2001), and have higher error rates than monolinguals, even when naming pictures in their dominant language (Gollan et al., 2005; Gollan & Acenas, 2004; Ivanova & Costa, 2008; Roberts et al., 2002). Using directly translated versions of tests designed in a different language (e.g., the Boston Naming Test) may magnify differences between bilinguals and monolinguals and provide an inaccurate assessment of language proficiency and dominance (Gollan et al., 2012).

Attempts to create linguistically equivalent translations may reduce, but not eliminate, this negative effect on performance depending on how equivalence is judged given idiosyncratic differences between languages that cannot be easily controlled, such as cultural interpretation, differences in word frequency and length, dialectal variation, and familiarity with the terms and/or concepts (Peña, 2007; Restrepo & Silverman, 2001).

Despite evidence that linguistic factors can reduce test sensitivity, little research has addressed the possible effects of language of testing on word list learning and memory tests. Word list learning tasks generate some of the most widely used measures for detection of preclinical AD (Weissberger et al., 2017; for review see Belleville et al., 2017), including the amount of learning across repeated presentation—immediate recall trials (i.e., hearing/reading the list and then immediately producing the words), delayed recall (i.e., producing the words after an interposed time interval filled with an unrelated cognitive task), and delayed recognition (i.e., recognizing the words after an interposed time interval filled with an unrelated cognitive task; Bondi et al., 1994). Additional features such as susceptibility to proactive interference or a reduction in semantic clustering are known to be sensitive to AD (Delis et al., 2010; Loewenstein et al., 2016; Rosselli et al., 2019) and can be derived from tests that include multiple word lists or words from multiple semantic categories (Breton et al., 2021; Delis et al., 1987; González et al., 2002; Rosselli et al., 2019).

Another aspect of performance on a word list learning test that is affected by AD is the well-established primacy effect – better immediate free recall of words from the beginning of a list compared to words positioned in the middle of the list (Atkinson & Shiffrin, 1968). The primacy effect is often attributed to greater attention to early list words and their more extensive rehearsal in long-term (episodic) memory. A recency effect also occurs with better recall for items at the end of the list than for middle items, an effect attributed to immediate production of words that are still in short-term memory at time of recall. Studies have demonstrated reduced primacy effects in individuals with mild cognitive impairment (MCI) and/or AD while recency effects remained unaffected (Bayley et al., 2000; Foldi et al., 2003; Howieson et al., 2011; Kloth et al., 2020; Moser et al., 2014). In addition, there is a negative association between primary effects (but not recency effects) and AD pathology (Gicas et al., 2020). A reduced primacy effect may also be a marker of preclinical or prodromal AD (Kloth et al., 2020). A reduced primacy effect on the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) Word List Learning Test improved prediction of future cognitive decline in individuals with MCI (Cunha et al., 2012; Moms et al., 1989). Clinically normal individuals with elevated cerebrospinal fluid levels of phosphorylated tau (a biomarker of AD) and a reduced primacy effect on the Buschke Selective Reminding Test (a 16-item word list memory and recall test) had increased risk of developing dementia related to AD, but only if they also had smaller hippocampal volume (Bruno et al., 2015).

Given the potential of the primacy effect as a cognitive marker of preclinical AD, it is important to understand the possible effects of language of testing on this feature of word list learning performance. Serial position effects (including primacy and recency effects) and their potential sensitivity to AD in Hispanics have not been examined. The present study examined serial position effects on the CERAD Word List Learning Test (Fillenbaum et al., 2007; Morris et al., 1989) in cognitively intact Hispanic decliners vs. non-decliners (i.e., controls) to determine if performance would be sensitive to future cognitive decline in this demographic. English and Spanish versions of the CERAD Word List Learning Test were used, as appropriate, which made it possible to examine language of testing effects. Previous studies have shown that English and Spanish versions of the CERAD Word List Learning Test have equivalent ability to detect memory impairment after controlling for education level (Carrion-Baralt et al., 2009; Fillenbaum et al., 2007; Ostrosky-Solis et al., 1999; but see O’Bryant et al., 2018); however, one study reported that the CERAD was less consistent in distinguishing levels of dementia severity in Spanish speaking than English-speaking participants (Fillenbaum et al., 2005). The CERAD Word List Learning Test may be well suited for examining serial position effects since it does not include semantically related words which could lead to clustering that counteracts serial position effects and is an optimal length for generating primacy effects. Previous studies in young monolingual participants have shown that primacy effects are maximized in free recall of shorter (6–8 words) compared to longer (14–15 words) word lists (Ward et al., 2010). The CERAD word list is 10 items (repeated over three learning trials) which should produce strong primacy effects.

Methods
Participants

The research protocol was approved by the UCSD Institutional Review Board in accordance with the Helsinki Declaration. Participants were recruited from the UCSD Alzheimer’s Disease Research Center (ADRC). Annual ADRC evaluations included detailed clinical and medical history, brief medical examination, neurological and neuropsychological assessment, screening for depression and other psychiatric symptoms, assessment of functional activities of daily living, and laboratory tests. At the completion of the evaluation two ADRC board-certified neurologists reviewed all information (in consultation with neuropsychologists) and classified each participant as either cognitively normal or cognitively impaired, and if cognitively impaired, diagnosed them with MCI (using current NIA-AA diagnostic criteria; Albert et al., 2011), AD (using current NIA-AA diagnostic criteria; McKhann et al., 2012), or another neurodegenerative disease (e.g., frontotemporal dementia, dementia with Lewy bodies) based on published criteria. Exclusion criteria were history of stroke, severe head trauma, substance abuse, or major neurological (other than AD dementia or MCI), psychiatric or metabolic disorder. All ADRC procedures received institutional...
Participant recruitment

Data from 57 Hispanic participants were included in analyses: 20 who were initially diagnosed as cognitively normal or MCI but had subsequent cognitive decline and eventually received a diagnosis of AD (decliners) and 37 who remained cognitively stable (controls) over at least 2 subsequent ADRC longitudinal evaluations. Of the 20 decliners, 11 had a diagnosis of MCI in the year prior to diagnosis, 3 demonstrated memory impairment without MCI, and 6 were cognitively normal but proceeded to decline the following year. Following Weissberger et al. (2013), participants were judged to be cognitively normal based on medical and neuropsychological assessments (see Galasko et al., 1994, for more details) and decliners were identified as individuals that converted from cognitively normal or MCI but had subsequent diagnosis of probable AD. Provable AD was determined by using criteria developed by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s Disease and Related Disorders Association (using current NIA-AA diagnostic criteria; McKhann et al., 2012).

Decliners and controls were matched for age, education, number of prior ADRC annual evaluations they had received, and did not differ significantly in Mini Mental State Exam (MMSE) scores which were in the normal range at the evaluation of interest (i.e., the annual evaluation completed in the year prior (or as close as possible to the year prior) to the year of diagnosis of AD. Decliners and controls were also equivalent in degree of bilingualism as measured by their relative ability to name pictures in each language on the Multilingual Naming Test (see Table 1). For this purpose, we calculated a Bilingual Index Score which divides the language with the lower naming score by the language with the higher naming score (a score of 1 represents a perfectly balanced bilingual who named the same number of pictures in each language; Gollan et al., 2010, 2012). Data from participants with mild and moderate AD from the same longitudinal study are shown in Tables 1 and 2 and in Figure 1 for visual comparison purposes only. These participants were not included in the analyses. Individuals with mild and moderate AD demonstrated a step-wise decline in overall recall and in the primary effect across severity of AD. For example, primacy scores were lowest in participants with moderate AD, followed by mild AD, then decliners, while controls exhibited robust primacy effects.

Materials and procedure

Participants were tested annually on the CERAD Word List Learning Test (Welsh et al., 1994) in English or Spanish as part of their annual neuropsychological evaluation in the UCSD ADRC longitudinal study. The test consists of immediate free recall of a
items recalled from the first three words on the presented list on Trial 1 (range 0–3). We focused on Trial 1 because primacy scores flatten with increasing learning trials (as more items from the middle of the list are recalled). Using the primacy score as the dependent variable, we examined the effects of group (decliner, control) and the language used for assessment (English, Spanish) in a 2 × 2 ANCOVA with group and language of testing as fixed factors, and age and education level as continuous covariates. For comparison, we also report analyses comparing decliners and controls on total correct, delayed recall correct, and five other neuropsychological (non-memory) tests (i.e., the Dementia Rating Scale (DRS), Trail-Making Test B, Semantic Fluency, Letter Fluency, and dominant language MINT score). Data were analyzed using SPSS, version 26 (IBM; Chicago, IL). All data have been uploaded to Open Science Framework (https://osf.io/hgm3e/).

Results

Performance on measures from the CERAD Word List Learning Test is shown in Table 2. Analysis of primacy scores, controlling for age and education, revealed a main effect of group F(1,58) = 5.97; \( \eta_p^2 = .11; p = .02 \), and no main effect of language F(1,58) = 1.47; \( \eta_p^2 = .03; p = .23 \), but a significant interaction F(1,58) = 4.54; \( \eta_p^2 = .08; p = .04 \), showing that the difference between decliners and controls was driven by participants tested in Spanish, see Tables A1–A2, Figures 2–3 (controls tested in Spanish recalled 1.7 of the 3 primacy words on average, while decliners recalled 0.69 on average). The effects of age and education were not significant, all ps ≥ .11.

Analysis of total correct showed a main effect of group F(1,58) = 11.99; \( \eta_p^2 = .19; p = .001 \), with decliners having lower total correct than controls. There is also a main effect of language, F(1,58) = 5.02; \( \eta_p^2 = .09; p = .03 \), with lower total correct for those tested in Spanish versus in English. The interaction between language and group was not significant, F(1,58) = 2.64; \( \eta_p^2 = .06; p = .11 \). In this ANCOVA, age was marginally significant, F(1,58) = 3.04; \( \eta_p^2 = .06; p = .09 \), and the effect of education was not significant, \( F < 1 \).

Analysis of delayed recall revealed a main effect of group, F(1,58) = 10.96; \( \eta_p^2 = .18; p = .002 \), with decliners recalling less words than controls. The main effect of language was marginally significant, F(1,58) = 3.66; \( \eta_p^2 = .07; p = .06 \), as was the interaction between language and group, F(1,58) = 3.53; \( \eta_p^2 = .06; p = .07 \). Similar to total recall, the covariate of age was marginally significant, F(1,58) = 3.67; \( \eta_p^2 = .07; p = .06 \), and the effect of education was not significant, \( F < 1 \).

All analyses comparing decliners and controls on five non-memory tests (i.e., the Dementia Rating Scale (DRS), Trail-Making Test B, Semantic Fluency, Letter Fluency, and dominant language MINT score; see Table 1) revealed main effects of group (all ps ≤ .02), no main effects of language (all ps ≥ .38), and no significant interactions (ps ≥ .11)\(^1\). Three tests produced main effects of education (DRS, \( p < .001 \); Trail-Making Test B, \( p = .001 \); dominant language MINT score, \( p = .04 \)), and two produced significant age effects (DRS, \( p = .03 \); Trail-Making Test B, \( p = .02 \)). These findings differed from those for the CERAD Word List Learning Test; specifically, the interaction between group and language was significant, \( F(1,58) = 5.22; \eta_p^2 = .08; p = .03 \).

\(^1\) All ps ≥ .25 except for the Dementia Rating Scale which exhibited a trend (p = .11) toward an interaction in the opposite direction as the CERAD measures, with the difference between decliners and controls being greater for participants tested in English (controls scored 135.4 on average, while decliners scored 127.3, a difference of 8.1) than for participants tested in Spanish (controls scored 132.3 on average, while decliners scored 128.4, a difference of just 3.9 points).
The results of the present study show that a number of measures from the CERAD Word List Learning Test, including the primacy score, were worse in decliners than in controls in the year prior to receiving a diagnosis of AD. Decliners scored worse than controls on total words recalled during learning and delayed recall consistent with previous studies that have shown subtle episodic memory decline in individuals with prodromal/preclinical AD (Twamley et al., 2006). Of particular interest, decliners exhibited a reduced primacy effect relative to their previous years of testing (within participant), and by the time of diagnosis (the year subsequent to the year shown in Tables 1–2 and in Figure 1) their primacy scores were at the same level as those tested in Spanish (see Figure 2). This provides converging evidence for the vulnerability of primacy scores to preclinical AD but leaves open the question of why this effect occurred 1–2 years sooner in those tested in Spanish. Additional information is provided in the Appendix which shows participant demographics (Table A1), test scores (Table A2), serial position curves (Figure A1), and box plots (Figure A2) broken down by language of testing.

One possible difference that might explain the more prominent drop in the primacy effect in participants tested in Spanish rather than English might be that those tested in English completed a slightly different and longer battery of cognitive tests as part of their annual longitudinal evaluation than those tested in Spanish, potentially leading to more fatigue, and only those tested in English also completed another word list learning test (the California Verbal Learning Test) prior to the CERAD Word List Learning Test (the word list learning tests were separated by 1–2 hours filled with unrelated testing and no words overlapped on the two tests). Both groups also remembered short non-overlapping word lists for the MMSE. Another difference is that participants tested in Spanish were more likely to have a diagnosis of MCI in the year prior to diagnosis (8/11 or 73%) than were those tested in English (3/9 or 33%). This might reflect greater difficulty diagnosing AD in participants tested in Spanish (due to possible cultural, linguistic, and demographic factors) so that they were further along in their decline trajectory than those tested in English. Consistent with this possibility, decliners tested in Spanish tended to have numerically lower test scores than decliners tested in English, though none of these differences were significant, all $p \geq 0.09$. However, this cannot provide a complete explanation for the language of testing differences we observed since the sharp drop in primacy scores for participants tested in Spanish was driven at least in part by unusually high primacy scores for controls tested in Spanish (see Table A2).

To further explore the effects of language of testing on CERAD Word List Test primacy effects, Figure 3 illustrates the proportion of participants in each group that recalled each of the first three words on Trial 1. This showed that controls tested in Spanish tended to recall primacy items more often than did controls tested in English (even though controls tested in Spanish had lower but not in late list positions for decliners relative to controls. This finding replicates studies that have demonstrated the sensitivity of primacy effects to preclinical or prodromal AD (Bruno et al., 2015; Cunha et al., 2012), and extends the finding to Hispanic participants. This increases the generalizability of the result to diverse populations and boosts confidence in the utility of reduced primacy as a measure for early detection of AD.
Additional studies will be needed to determine if and why the nature of the primacy effect varied across languages. Our results raise the possibility that certain linguistic variables may heighten sensitivity of primary effects to preclinical AD and invite further study of the primacy effect in diverse populations. This effort seems worthwhile since the population of Hispanic elders is expected to grow seven-fold by 2050 (U.S. Census Bureau, 2004), the prevalence of AD is expected to increase exponentially (Novak & Riggs, 2004), and understanding linguistic differences will be broadly informative as to why the primacy effect is sensitive to early AD.

Funding statement. This research was supported by grants from NIA [R01 (AG076415), PS0 (AG05131), P30 (AG062429)] and NSF (BCS 1923065). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NIH or NSF.

Competing interests. We have no conflicts of interest to disclose.

References


Figure 4. Primacy score on Trial 1 of the CERAD plotted against DRS score for all participants tested in each language including controls, decliners, mild AD, and moderate AD (see Table 1).

education level than controls tested in English). Thus, greater sensitivity of the primary effect to preclinical AD in Spanish than in English seemed to be driven as much by performance of the controls as by performance of the decliners. It might seem that this could be attributed to differences in word length (Spanish words tend to be longer than English words) but note that in both languages the first word was twice as long as the second and third words (in number of syllables). As Figure 3 shows, primary effects were evident in controls in both languages even within the first three words on the list – they recalled the first word most often and the third word least often. However, the first and best recalled word in each language was also the longest of the three words confounding serial position and word length (note that longer words are recalled better than short words when long and short words are intermixed within a list; Katkov et al., 2014). The greater sensitivity of the primary effect in Spanish than in English was limited to the preclinical AD stage. When all participants were included in the analysis, including those with mild and moderate dementia, primary scores were significantly correlated with scores on the DRS and these correlations were equally strong in both languages, or if anything, were stronger in English, \( r = .415, p < .001 \) than in Spanish, \( r = .381, p = .006 \) (see Figure 4).

The present results also showed that a number of neuropsychological tests besides the CERAD Word List Learning Test are sensitive to preclinical AD in elderly Hispanic individuals, and about equally so for those who prefer to be tested in English versus Spanish. Despite similar MMSE scores, decliners performed worse than controls on the Trail-Making Test B, a picture naming test (here measured with the MINT instead of the BNT as in Weissberger et al., 2013), the DRS, and semantic and letter fluency tests. In contrast to the present results, the DRS and fluency tests did not differ between Hispanic decliners and controls in Weissberger et al. (2013), a difference that might be attributable to differences in the interval between testing and conversion to probable AD: under 2 years in the present study (see Table 1) and closer to 5 years in Weissberger et al. (2013). That is, decliners in the present study were likely further along in the course of AD than those in Weissberger et al. (2013).
Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology, 24*(3), 402–411.


Appendix Data split by language of CERAD administration

### Table A1. Characteristics of cognitively healthy controls and demographically matched decliners split by language

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th>Spanish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n = 21)</td>
<td>Decliner (n = 9)</td>
<td>Control (n = 16)</td>
<td>Decliner (n = 11)</td>
</tr>
<tr>
<td>Age</td>
<td>M = 79.9, SD = 7.1</td>
<td>M = 83.2, SD = 6.7</td>
<td>M = 75.8, SD = 7.1</td>
<td>M = 76.3, SD = 8.7</td>
</tr>
<tr>
<td>Education</td>
<td>M = 14.8, SD = 2.6</td>
<td>M = 13.8, SD = 3.2</td>
<td>M = 10.4, SD = 3.0</td>
<td>M = 9.8, SD = 2.3</td>
</tr>
<tr>
<td>Years to conversion</td>
<td>. .</td>
<td>M = −1.2, SD = 4.4</td>
<td>. .</td>
<td>M = −1.5, SD = 1.2</td>
</tr>
<tr>
<td>Number of previous evaluations</td>
<td>M = 8.0, SD = 5.1</td>
<td>M = 10.6, SD = 6.0</td>
<td>M = 9.8, SD = 7.4</td>
<td>M = 7.6, SD = 6.0</td>
</tr>
<tr>
<td>DRS</td>
<td>M = 137.3, SD = 3.9</td>
<td>M = 127.6, SD = 5.1</td>
<td>M = 130.9, SD = 7.1</td>
<td>M = 126.5, SD = 5.6</td>
</tr>
<tr>
<td>Trail-Making Test, Part A</td>
<td>M = 28.5, SD = 1.9</td>
<td>M = 27.3, SD = 2.5</td>
<td>M = 26.4, SD = 7.5</td>
<td>M = 26.7, SD = 2.2</td>
</tr>
<tr>
<td>Trail-Making Test, Part B</td>
<td>M = 126.5, SD = 65.8</td>
<td>M = 241.8, SD = 72.6</td>
<td>M = 171.3, SD = 89.0</td>
<td>M = 229.9, SD = 67.1</td>
</tr>
<tr>
<td>Semantic Fluency</td>
<td>M = 42.0, SD = 10.0</td>
<td>M = 33.4, SD = 10.7</td>
<td>M = 40.2, SD = 12.3</td>
<td>M = 32.5, SD = 3.7</td>
</tr>
<tr>
<td>Letter Fluency</td>
<td>M = 35.0, SD = 12.0</td>
<td>M = 27.4, SD = 12.1</td>
<td>M = 37.3, SD = 8.7</td>
<td>M = 28.5, SD = 6.3</td>
</tr>
<tr>
<td>Digit Span Forward</td>
<td>M = 6.0, SD = 1.1</td>
<td>M = 6.1, SD = 1.1</td>
<td>M = 4.8, SD = 7.</td>
<td>M = 5.4, SD = 5</td>
</tr>
<tr>
<td>Digit Span Backward</td>
<td>M = 4.1, SD = 1.1</td>
<td>M = 3.7, SD = 1.3</td>
<td>M = 3.9, SD = 1.0</td>
<td>M = 3.6, SD = 5</td>
</tr>
<tr>
<td>MINT dominant language</td>
<td>M = 64.8, SD = 7.1</td>
<td>M = 58.9, SD = 6.5</td>
<td>M = 62.3, SD = 2.2</td>
<td>M = 59.8, SD = 3.2</td>
</tr>
<tr>
<td>MINT Nondominant language</td>
<td>M = 43.5, SD = 17.5</td>
<td>M = 49.3, SD = 5.4</td>
<td>M = 44.0, SD = 12.8</td>
<td>M = 40.6, SD = 17.1</td>
</tr>
</tbody>
</table>

### Table A2. Performance of cognitively healthy controls and demographically matched decliners split by language

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th>Spanish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n = 21)</td>
<td>Decliner (n = 9)</td>
<td>Control (n = 16)</td>
<td>Decliner (n = 11)</td>
</tr>
<tr>
<td>Trial 1 correct</td>
<td>M = 4.5, SD = 1.6</td>
<td>M = 4.2, SD = 1.4</td>
<td>M = 4.8**, SD = 1.8</td>
<td>M = 2.9, SD = 1.1</td>
</tr>
<tr>
<td>Trial 2 correct</td>
<td>M = 7.0, SD = 1.5</td>
<td>M = 6.2, SD = 1.9</td>
<td>M = 7.0**, SD = 1.6</td>
<td>M = 5.1, SD = 1.4</td>
</tr>
<tr>
<td>Trial 3 correct</td>
<td>M = 7.8†, SD = 1.3</td>
<td>M = 6.9, SD = 1.2</td>
<td>M = 7.4**, SD = 1.0</td>
<td>M = 6.5, SD = 1.3</td>
</tr>
<tr>
<td>Total correct</td>
<td>M = 19.2, SD = 3.4</td>
<td>M = 17.3, SD = 3.8</td>
<td>M = 19.2***, SD = 3.4</td>
<td>M = 14.5, SD = 2.2</td>
</tr>
<tr>
<td>Delay correct</td>
<td>M = 6.1, SD = 1.9</td>
<td>M = 5.2, SD = 2.5</td>
<td>M = 6.3***, SD = 1.5</td>
<td>M = 3.6, SD = 1.5</td>
</tr>
<tr>
<td>Total intrusions trials 1-3</td>
<td>M = 0.6*, SD = 1.0</td>
<td>M = 2.0, SD = 2.2</td>
<td>M = 0.8, SD = 1.2</td>
<td>M = 1.1, SD = 1.6</td>
</tr>
<tr>
<td>Delay intrusions</td>
<td>M = 0.2†, SD = 0.7</td>
<td>M = 0.9, SD = 1.2</td>
<td>M = 0.4, SD = 0.5</td>
<td>M = 0.4, SD = 0.8</td>
</tr>
<tr>
<td>Trial 1 primacy</td>
<td>M = 1.4, SD = 1.0</td>
<td>M = 1.3, SD = 0.9</td>
<td>M = 1.9***, SD = 0.7</td>
<td>M = 0.9, SD = 0.7</td>
</tr>
<tr>
<td>Trial 1 middle</td>
<td>M = 1.4, SD = 1.0</td>
<td>M = 1.1, SD = 0.9</td>
<td>M = 1.1, SD = 1.2</td>
<td>M = 0.6, SD = 0.8</td>
</tr>
<tr>
<td>Trial 1 recency</td>
<td>M = 1.6, SD = 0.8</td>
<td>M = 1.8, SD = 1.2</td>
<td>M = 1.7, SD = 0.7</td>
<td>M = 1.4, SD = 0.7</td>
</tr>
<tr>
<td>Recognition correct YES</td>
<td>M = 9.5, SD = 0.9</td>
<td>M = 9.4, SD = 0.7</td>
<td>M = 9.7***, SD = 0.6</td>
<td>M = 8.7, SD = 1.1</td>
</tr>
<tr>
<td>Recognition correct NO</td>
<td>M = 9.8, SD = 0.4</td>
<td>M = 9.9, SD = 0.3</td>
<td>M = 9.8†, SD = 0.4</td>
<td>M = 10.0, SD = 0</td>
</tr>
<tr>
<td>Bilingual Index Score</td>
<td>M = 0.7, SD = 0.3</td>
<td>M = 0.8, SD = 0.2</td>
<td>M = 0.7, SD = 0.2</td>
<td>M = 0.7, SD = 0.3</td>
</tr>
</tbody>
</table>

**p ≤ 0.001, *p ≤ 0.01, †p < 0.05, ‡p ≤ 0.10; p-values comparing decliners to controls within language.

Note: All group comparisons should be interpreted with caution given group sizes.

Note: The total possible correct for primacy (list items 1-3) and recency scores (list items 8-10) is 3 while the total possible correct for the middle score is 4 (list items 4-7).
Figure A1. Trial 1 serial position curves for each participant group split by language.

Figure A2. Trial 1 box plots split by group and language.