Published in final edited form as:

J Cogn Neurosci. 2012 May ; 24(5): 1205–1223. doi:10.1162/jocn_a_00143.

The Influence of Language Proficiency on Lexical Semantic Processing in Native and Late Learners of English

Aaron J. Newman[1], **Antoine Tremblay**[2,*], **Emily S. Nichols**[1], **Helen J. Neville**[3], and **Michael T. Ullman**[2]

[1]Dalhousie University

[2]Georgetown University

[3]University of Oregon

Abstract

We investigated the influence of English proficiency on ERPs elicited by lexical semantic violations in English sentences, in both native English speakers and native Spanish speakers who learned English in adulthood. All participants were administered a standardized test of English proficiency, and data were analyzed using linear mixed effects (LME) modeling. Relative to native learners, late learners showed reduced amplitude and delayed onset of the N400 component associated with reading semantic violations. As well, after the N400 late learners showed reduced anterior negative scalp potentials and increased posterior potentials. In both native and late learners, N400 amplitudes to semantically appropriate words were larger for people with lower English proficiency. N400 amplitudes to semantic violations, however, were not influenced by proficiency. Although both N400 onset latency and the late ERP effects differed between L1 and L2 learners, neither correlated with proficiency. Different approaches to dealing with the high degree of correlation between proficiency and native/late learner group status are discussed in the context of LME modeling. The results thus indicate that proficiency can modulate ERP effects in both L1 and L2 learners, and for some measures (in this case, N400 amplitude), L1–L2 differences may be entirely accounted for by proficiency. On the other hand, not all effects of L2 learning can be attributed to proficiency. Rather, the differences in N400 onset and the post-N400 violation effects appear to reflect fundamental differences in L1–L2 processing.

INTRODUCTION

With increasing age of acquisition (AoA) of a second language (L2), evidence suggests decreasing ultimate achievement and increasing variance of proficiency among individuals (Birdsong & Molis, 2001; Flege, Yeni-Komshian, & Liu, 1999; Johnson & Newport, 1989). The reasons for these changes in language learning ability are not well understood and likely involve a combination of maturational changes, differences in social and learning

Reprint requests should be sent to Aaron J. Newman, Department of Psychology, Life Sciences Centre, Dalhousie University, Halifax, NS, Canada, B3H 4R2, or via Aaron.Newman@dal.ca.

*Antoine Tremblay is now at the Issak Walton Killam Health Centre.

environments, and the cumulative effects of learning and use of an L1 over time (Morgan-Short, Sanz, & Ullman, 2010; Weber-Fox & Neville, 1996; Johnson & Newport, 1989; Lenneberg, 1967). One approach to better understanding the effects of age on L2 abilities is the use of neuroimaging techniques such as ERPs and fMRI. Studies have typically compared highly proficient L1 speakers with L2 speakers of varying (and generally lower) proficiency. Because L1/L2 status and proficiency are confounded, it is often unclear whether any observed L1–L2 differences are attributable to L2s being processed in qualitatively different ways from L1s or simply because of differences in the amount of effort required for language processing. The answer to this question fundamentally impacts how we interpret neuroimaging studies of L2 learners. One approach to this question is to compare L1 and L2 learners of comparable proficiency, but differing AoA. However, because most L2 learners do not achieve native levels of proficiency, such an approach is restricted either to a narrow sample of L2 learners, or compares higher-than-average proficiency L2 learners with lower-than-average L1 learners. An additional concern is that most studies seem to treat L1 learners as having universally maximal proficiency, when in fact their scores on vocabulary and grammar tests vary (Hammill, Brown, Larsen, & Wiederholt, 1994). In this study, we addressed these issues directly by measuring both brain activation using ERPs and English proficiency in both L1 and L2 learners, treating proficiency as a continuous variable.

Several studies have previously demonstrated effects of proficiency in L2 learners. Some areas show increased activation among more proficient learners. Using fMRI, Wartenburger et al. (2003) found that high-proficiency late learners showed greater activation than lower-proficiency late learners in temporo-parietal regions when performing a grammatical judgment task in their L2, thus suggesting that activation in this area is sensitive to proficiency when AoA is controlled. Newman-Norlund, Frey, Petitto, and Grafton (2006) found an effect of proficiency in the left inferior frontal gyrus (LIFG) activation in a longitudinal study of artificial grammar learning, with activation during sentence processing increasing as mastery of the language increased. Conversely, Meschyan and Hernandez (2006) found greater activation of components of the articulatory motor system when Spanish–English bilinguals read words in their less proficient L1 than in their more proficient L2.

Not all differences in brain activation between L1 and L2 learners can be attributed to proficiency, however. For example, when late L2 learners were compared with early L2 learners of comparable high proficiency, greater activation was found in the LIFG—an effect of later acquisition where the possible effect of proficiency was controlled for (Wartenburger et al., 2003). Similarly, Perani et al. (2003) found increased LIFG activity for a phonological fluency task performed in L2 versus L1, in early L2 learners who were assumed to have relatively high proficiency (although proficiency was not explicitly measured). These studies provide evidence that separable effects of proficiency and AoA may be detected using neuroimaging.

Several ERP studies have also investigated the relationship between proficiency and brain activation in L2 learners. These studies generally report more native-like patterns of scalp activity in more proficient L2 learners. The ERP components that have been most studied

are the following: the N400, a negativity peaking around 400 msec postword onset, which is larger for semantically anomalous than congruent words; the P600, a positivity that is typically larger in response to syntactic violations; and a LAN around 150–500 msec, which is also sensitive to syntactic congruity. In studies of both lexical semantic and syntactic processing, less proficient learners show delayed onsets and/or peaks of components, reduced amplitudes, and in some cases qualitatively different or even absent components (Midgley, Holcomb, & Grainger, 2009; Hahne, Mueller, & Clahsen, 2006; Rossi, Gugler, Friederici, & Hahne, 2006; Elston-Güttler, Paulmann, & Kotz, 2005; Moreno & Kutas, 2005; Ojima, Nakata, & Kakigi, 2005; Kotz & Elston-Güttler, 2004; Phillips, Segalowitz, Brien, & Yamasaki, 2004). Similar results have been found in studies of artificial or miniature languages where the language exposure is known and controlled (Morgan-Short et al., 2010; Mueller, Oberecker, & Friederici, 2009; Mueller, Hahne, Fujii, & Friederici, 2005; Friederici, Steinhauer, & Pfeifer, 2002).

Taken together, these data suggest that it is critical to account for proficiency in neuroimaging studies of L2 learners. One methodological issue, though, is that “high” and “low” proficiency are often defined fairly arbitrarily, such as a median split of a proficiency measure or on self-reported amounts of usage. Such an approach makes it hard to compare one group of “high-proficiency” learners to another. Additionally, the practice of dichotomizing continuous variables such as proficiency leads to a loss of power and reduced effect sizes as well as increasing the likelihood of finding spurious significant effects (Cohen, 1983). A preferable approach is to use a standardized test of proficiency and treat scores as falling along a continuum.

A second question that has not been well addressed in the literature is how proficiency affects neurocognitive measures of language processing in L1 learners. L1 proficiency is rarely measured in lieu of the implicit, but erroneous, assumption that all L1 speakers perform at ceiling. It is crucial to determine whether the variation in brain activation associated with proficiency is similar in L1 or L2 learners or qualitatively different. If proficiency modulates neural activation in the same way in L1 and L2 learners, then we gain insight into the associated processes, but not into the question of why L2 learners generally have lower proficiency. Conversely, if differences remain after controlling for the effects of proficiency, then we can be assured that we are looking at the effects of L1/L2 learner status.

Two studies have investigated the effects of L1 proficiency on ERPs. Pakulak and Neville (2010) found earlier-latency anterior negativities and higher-amplitude P600 components in higher- than lower-proficiency native speakers of English, in response to syntactic phrase structure violations. Weber-Fox, Davis, and Cuadrado (2003) compared “high” and “normal” proficiency L1 English learners’ responses to felicitous and semantically anomalous words during sentence processing. Late (400–600 msec) negative responses to semantically congruous open class (i.e., content) words were reduced over posterior electrodes but enhanced over anterior ones, in the high proficiency group. As well, responses to semantically anomalous words were reduced in high-proficiency learners, and the amplitude of this effect correlated with the standardized measure of proficiency used (the TOAL-3; Hammill et al., 1994; also used by Pakulak & Neville, 2010).

One study of L2 speakers did treat proficiency as a continuous variable. Moreno and Kutas (2005) found that a measure of vocabulary knowledge correlated with the timing, but not the amplitude, of the N400 elicited by semantic violations in both L1 and L2 learners. N400 peak latencies over a left posterior electrode were earlier in people with higher vocabulary scores and in participants' dominant language (regardless of whether this was their L1 or L2). N400 peak latency also increased with AoA in this group, but using stepwise linear regression Moreno and Kutas showed that whichever variable (proficiency or AoA) was entered first, the second explained additional variance. These results suggest separable effects of AoA and proficiency.

The evidence thus indicates that proficiency is an important factor affecting brain activation in L2 and even in L1. In L2, effects of proficiency and AoA may be separable, and so it is important to include both as predictors. Furthermore, it is important to determine whether proficiency affects patterns of brain activation similarly in L1 and L2. By taking this approach, we have the power to determine whether any observed differences between L1 and L2 learners are simply because of the groups' falling, on average, at different points along the continuum of proficiency, or if the differences can be attributed to differences in how L1 and L2 are processed.

The goals of this study were to (1) characterize effects of proficiency on brain activation during the processing of semantically congruous and incongruous sentences, in both L1 and L2, treating proficiency as a continuous measure, and (2) separate the effects of proficiency on lexical semantic processing from those of L1 versus late L2 learner status. Previous studies have taken important first steps into exploring the relationship between proficiency and brain activation. In this study, we aimed to make several further advances. For one, we used a standardized test of proficiency (the TOAL-3) that does not show ceiling effects in native speakers and tested both L1 and late L2 learners. As well, we used linear mixed effects (LME) modeling so that proficiency could be treated as a continuous variable included alongside numerous factorial variables (e.g., condition; electrode position). This improves on previous approaches that have either dichotomized proficiency to incorporate it into an ANOVA framework (at the cost of statistical power/sensitivity), or used simple linear regression (at the expense of considering numerous predictive factors in a single analysis, e.g., limiting the regression to a single electrode). Another issue that has not typically been addressed is the fact that proficiency and L1/L2 learner status are typically highly collinear. Moreno and Kutas (2005) took one approach in using stepwise linear regression to test whether, after one of these variables had been entered, the addition of the other explained additional variance. We built upon this foundation by using both the stepwise approach, and an alternative in which the variance in proficiency because of group was first removed and then the residual variance included in the LME models alongside group, condition, and electrode factors. In doing so we were further able to assess the robustness of any effects that were found.

L2 learners in this study were native Spanish speakers with a mean age of first exposure to English of 10 years, mean age of first arrival in an English-speaking country of 24 years, and an average of 8 years living in an English-speaking country. Most participants reported that, although they had childhood exposure to English, typically through school and/or

television, they did not feel they had achieved any significant level of fluency until moving to an English-speaking country in adulthood. We measured vocabulary and grammatical proficiency in both native and late learners of English, using the TOAL-3. This is a standardized test battery that assesses both grammar and vocabulary skills, has norms up to age of 24, does not typically show ceiling effects even among English L1 speakers, and has been used in previous ERP studies of language (Pakulak & Neville, 2010; Weber-Fox et al., 2003). We used a paradigm involving lexical semantic violations (e.g., *The Irishman sipped Todd's* thunder *at the party*), which typically elicit an N400 relative to well-formed control sentences. The N400 is thought to reflect aspects of lexical access and the postlexical integration of word meanings into episodic memory (Lau, Phillips, & Poeppel, 2008; Kutas & Federmeier, 2000).

We predicted that learner status (native or late) and proficiency would have separable effects on ERPs. Specifically, following previous studies, we predicted that the latency and amplitude of the N400 elicited by semantic anomalies would be later and smaller, respectively, in late learners. However, we further predicted, on the basis of Moreno and Kutas' (2005) data, that the differences in latency and possibly in amplitude would largely be accounted for by proficiency, controlling for L1/L2 learner status. Thus, we predicted that a similar relationship between proficiency and N400 latency and amplitude would be found for L1 and L2 learners of English. To the extent that this prediction did not hold, and group differences were observed once proficiency had been accounted for in the model, we would attribute such group differences as likely stemming from an effect of late L2 acquisition independent of proficiency. An additional possibility was that while proficiency might modulate one or more properties of the N400 in both groups, it might do so in different ways between groups. Such an interaction could be interpreted as evidence that, although proficiency modulates lexical processing, it does so in different ways, depending on whether the language is acquired from birth or in early adulthood as an L2.

METHODS

Participants

Nineteen native English speakers (mean age = 23.3 years, *SD* = 7.1 years, range = 18–51 years; mean years of education = 14.4 years, *SD* = 1.7 years) and 19 native Spanish speakers (mean age = 34 years, *SD* = 6.6 years, range = 21–46 years; mean years of education = 17.9 years, *SD* = 3.0 years) took part in this study. The L2 learners were thus older, $R = 0.33$, $F(1, 33) = 17.5$, $p = .0002$, and had more years of education, $R = 0.31$, $F(1, 33) = 16.6$, $p = .0003$. These differences in age and education level were taken into account in the analyses as described below. All participants were men, right-handed, and without any reported neurological or psychiatric pathology. Native Spanish speakers' mean age of first exposure to English was 9.6 years (*SD* = 9.6 years; from 0 to 24 years). First exposure was typically in school, taught by a nonnative English speaker, with 1–6 hr/week of formal instruction. The mean age at which native Spanish speakers moved to an English-speaking country where they were immersed in English was 24.2 years (*SD* = 6.9 years, range = 18–40 years) and the average length of time they had lived in an English-speaking country was 8.1 years (*SD* = 5.2 years, range = 1–19 years). All but 3 of the native English speakers reported some

knowledge of at least one other language. Participants were paid $20 for their participation. Study procedures were reviewed by the Georgetown University Institutional Review Board.

Materials

The target stimuli for this experiment consisted 64 simple declarative English sentences. Two versions of each sentence were created, one that was semantically acceptable (e.g., *The Irishman sipped Todd's whiskey at the party*) and the other in which the direct object of the verb was replaced with a noun, matched in lexical frequency, that did not make contextual sense (e.g., *The Irishman sipped Todd's thunder at the party*). Stimuli were counterbalanced across participants, such that each participant saw only the control or the anomalous version of a given sentence. An additional 192 sentences were used, including 32 sentences each with violations of regular past tense morphology, irregular past tense morphology, and syntactic phrase structure. The remaining sentences were grammatically and semantically acceptable. The complete set of stimuli are available in Newman, Ullman, Pancheva, Waligura, and Neville (2007). Because the present article is focused on the relationship between language proficiency and lexical semantic processing, the results of the grammatical violations will not be discussed here.

Participants were administered a general health screening and a language history questionnaire that included self-ratings of proficiency in each language known (on a 5-point Likert scale). The following subtests from the Test of Adult and Adolescent Language, third edition (Hammill et al., 1994), were also administered: Reading and Listening Vocabulary, and Listening, Reading, and Speaking Grammar.

ERP Recording and Preprocessing

Continuous EEG data were recorded from each participant via 64 tin electrodes sewn into a tight-fitting cap (Electro-Cap, Eaton, OH), referenced on-line to an electrode on the right mastoid bone (later rereferenced to the average of the left and right mastoid locations). Electrode positions were specified by the International 10–20 system (FP1/2, FP3/4, FPz, F1/2, F3/4, F5/6, F7/8, Fz, FF1/2, FF3/4, FC1/2, FC3/4, FC5/6, FC6/7, C1/2, C3/4, C5/6, Cz, T3/4, T5/6, CP1/2, CP3/4, CT5/6, CT7/8, P1/2, P3/4, P5/6, Pz, PO3/4, POz, O1/2, TO1/2, Oz, IN3/4, INz, left/right mastoid). EOG was recorded from electrodes positioned on the outer canthi of each eye as well as one electrode placed below the left eye. EEG was amplified (SAI model GTU-96/128BA; San Diego, CA) using a 3-dB cutoff, bandpass filtered 0.01–125 Hz and digitized at 256 Hz for recording on a desktop computer.

Trials with blinks, eye movements, or excessive noise were identified off-line (using a maximum peak-to-peak amplitude threshold tailored to each participant's data) and were discarded, as were trials containing blocking (defined as 10 or more time points having the same value). Data were digitally notch-filtered at 60 Hz. Trials to which participants responded correctly were averaged within each condition over an epoch of 200 msec prestimulus to 1500 msec poststimulus onset.

Procedure

After giving informed consent, participants completed the questionnaires and were administered the TOAL-3 sub-tests. The EEG cap was then applied, and impedances were lowered to <5 kΩ. Participants were then seated in a dimly lit, sound-attenuating booth 135 cm from a CRT monitor; stimulus words subtended 0.5° vertically and 1°–3° deg horizontally. Participants were given a response button box to hold in both hands. Each sentence was initiated by a button press from the participant and began with the outline of a box (7° × 3° visual angle) appearing on the computer monitor for a random period of 300–1100 msec. The words of the sentence were then presented one at a time, with each word displayed for 300 msec and a 200-msec delay between words. The outline of the box remained for 1500 msec after the last word of the sentence and was then replaced by a response prompt, “Good or bad?”, displayed on the screen. This remained visible until the participant responded (response buttons were counterbalanced across participants and across the four sets of stimuli), at which point a fixation cross was displayed until a button was pressed to initiate the next trial. Participants were given short breaks after every 50 sentences and could initiate breaks at any other time. Order of sentence presentation was randomized for each participant. Before the experimental stimuli being presented, participants performed a practice session consisting of 16 sentences, receiving feedback on their performance. Feedback was not provided for the experimental stimuli.

Statistical Analyses

To investigate the influence of proficiency and learner status on ERP amplitude, scalp distribution, and timing, we conducted LME modeling as implemented by the function *lmer()* from the *lme4* library in R version 2.10 (Bates, Maechler, & Bolker, 2009). LME is a relatively recent development in computational statistics based on restricted maximum likelihood estimation. Its use in the analysis of EEG data was advocated for by Bagiella, Sloan, and Heitjan (2000) and has been used by Pritchett et al. (2010), Wierda, van Rijn, Taatgen, and Martens (2010), Davidson and Indefrey (2007), and Moratti, Clementz, Gao, Ortiz, and Keil (2007).

LME models are a form of general linear model that include both fixed effects parameters and random effects. LME models offer several advantages over traditional repeated measures ANOVA that made such an approach desirable for the present ERP data (and for many other typical ERP data sets as well). For one, LME allows richer modeling of random effects (variables with levels that represent a random, nonreproducible, sample of a population such as participants or items) including multiple, crossed, and/or nested random effects. This can increase the accuracy and generalizability of the parameter estimate. The more complex random effects structure in our LME models enabled us to better account for the correlation in the residuals than traditional repeated measures AN(C)OVA models do and thus better approximate the assumption of uncorrelated model residuals. Furthermore, as in many ERP studies, our data were unbalanced. Although the results of an AN(C)OVA performed on unbalanced data need to be interpreted with caution, LME models can appropriately deal with unbalanced data by weighting the contribution of each group (e.g., participants in by-subject random intercepts and/or the violation and control levels of the condition factor in by-subject random adjustments for condition) according to the number of

observations in the group and the variation within and between the groups (Gelman & Hill, 2007, p. 254). Finally, LMEs properly deal with missing data and account for nonsphericity, common in ERP data, without the need for subsequent correction (e.g., Greenhouse–Geisser or Huynh–Feldt; Baayen, Davidson, & Bates, 2008; Bagiella et al., 2000).

In the analyses we conducted, predictors in the LME models included the fixed effects proficiency (TOAL-3 composite scores, centered by subtracting the mean score from each participant's score), group (native or late learners),[1] condition (control or violation), and electrode position. Electrodes were grouped into ROIs arranged in a 3 × 3 grid over the scalp (left/midline/right and anterior/ central/posterior); data from each electrode within an ROI were treated as repeated measures of that ROI. We included by-subject random adjustments for the intercept, condition, and ROI in the model specification as crossed random effects.[2] Because proficiency and group strongly correlated with one another ($R = 0.74$), we residualized proficiency with respect to group by taking the residuals of a linear model fit between proficiency and group (i.e., proficiency as a function of group; Tremblay & Tucker, 2011). The correlation between original and residualized proficiency scores was high ($R = 0.72$), indicating that residualized proficiency still captures the interindividual variability present in the raw proficiency scores.[3] After residualizing the scores, the probability density functions of the residualized native- and late-learner proficiency scores overlapped completely and were nonpredictive of group membership.

Identification of the optimal mixed-effects model was performed for each dependent measure through a series of iterative tests comparing progressively simpler models with more complex models using log-likelihood ratio testing (Tremblay, 2011; Tremblay & Tucker, 2011). This allows removal of interactions and variables that do not explain significant amounts of variance (Baayen et al., 2008). The optimal model was the one having the fewest factors and interactions that accounted for more variance than the next less complex model. Once the optimal model was obtained, outliers were removed, and the model was refitted (e.g., Tremblay & Tucker, 2011). After this refitting, the residuals were approximately normally distributed in all cases. The variance explained by each factor was examined by way of (sequential) *F* tests for main effects and interactions and *t* tests for specific contrasts. Exact determination of the denominator degrees of freedom (*df*) for LME models is difficult at best (Bates, 2005). Thus, we calculated both upper- and lower-bound values. These may be somewhat anticonservative and conservative, respectively, although when the number of data points is large (as in our data), these two values may in fact be very

[1]AoA was dichotomized as "group," rather than being treated as a continuous variable, because this variable had a bimodal distribution, with all native speakers having an AoA of 0. Such a distribution violates the assumption of normality.

[2]Theoetically speaking, every participant has violation and control values that differ, more or less, from other participants and the population means. By including a variable in both the fixed and random effects structure of a model, individual (by-subject) deviations from the fixed effects are estimated. The inclusion of by-subject adjustments for the two levels of the condition factor allowed us to model between-subject variability and obtain more accurate and generalizable estimates of these two levels as well as to properly deal with any imbalance in the data. The inclusion of ROI in the random effect structure enabled us to account for both individual variation in the scalp distribution of the effects, but also for individual spatial correlations between ROI levels. In each analysis performed, we assessed whether the inclusion of these random effects significantly improved the fit of the model to the data. To do this a model without individual adjustments for the violation and control levels of the condition factor and a model allowing for such adjustments were fitted and a log-likelihood ratio test was performed between these two models (Bagiella et al., 2000; Pinheiro & Bates, 2000). These tests proved significant, justifying the inclusion of these random effects.

[3]The correlation between (objective) proficiency and self-reported proficiency is .79. The correlation between residualized (objective) proficiency and self-reported proficiency is .37.

similar. *df* to compute upper-bound probability values were calculated as the number of data points minus the number of *df* used up by the fixed effects. Those for lower-bound *p* values were calculated as the number of data points minus the number of *df* used up by the fixed effects and the number of random effects in the model (i.e., *df* used up by individual adjustments for the two levels of the condition factor plus *df* used up for individual adjustments for the nine levels of the ROI factor). R package *LMERConvenienceFunctions* was used for the back-fitting of fixed effects, the forward-fitting of random effects, and the calculation of upper- and lower-bound *p* values (Tremblay, 2011).

RESULTS

Standardized Measures of Language Proficiency

The average scores for native learners were above the 60th percentile on all TOAL-3 subtests (based on norms from native American English-speaking adults aged 21–24 years), averaging in the 77th percentile (range of group average across tests: 62–92). In contrast, the late learners averaged in the 31st percentile (range of group average across tests: 13–47). Across the individual subtests, within the native learner group the percentile scores ranged from 5th to 98th, whereas within the late learner group percentile scores ranged from 1st to 98th. These are plotted in Figure 1A. A 2 (native vs. late learners) × 5 (TOAL subtests) mixed-effects ANOVA (function *aov()* in R v. 2.8.1; www.R-project.org) yielded significant main effects of Group, $F(1, 36) = 44.3$, $p < .001$, and Test, $F(4, 144) = 26.7$, $p < .001$, as well as a Group × Test interaction, $F(4, 144) = 3.13$, $p = .017$. Post hoc tests of the difference between groups for each TOAL subtest revealed significantly higher proficiency for native learners on each measure, all *p* values < .001.

We created composite scores for vocabulary and for grammar by summing the two vocabulary and three grammar subtests, respectively, for each participant. Across all participants, these were highly correlated with each other, $R = .76$, $p < .0001$. Thus, we chose to create a single measure of proficiency, TOAL-3 composite score, by summing the scores of the five individual subtests. TOAL-3 composite scores were significantly higher for native than for late learners, $F(1, 36) = 44.34$, $p < .0001$. Nevertheless, there was overlap between groups such that some late learners had higher TOAL-3 composite scores than some native speakers. This can be seen in Figure 1B, where the rank-ordered TOAL-3 composite scores of each participant are plotted.

Sentence Acceptability Judgments

Participants were quite accurate in discriminating correctly formed English sentences from those containing semantic violations. Native English speakers were 96.5% correct for control sentences and 94.1% correct in detecting violations. Late learners were 89.3% correct for control sentences and 80.1% correct for violation sentences. A mixed effects 2 (native vs. late learners) × 2 (control vs. violation sentences) ANOVA (using *aov()* in R) was performed using number of correct responses as the dependent variable. This yielded significant main effects of Group, $F(1, 36) = 19.3$, $p < .001$, and Sentence Type, $F(1, 36) = 9.4$, $p = .004$, as well as a significant Group × Violation interaction, $F(1, 36) = 6.5$, $p = .015$. Overall, native learners were more accurate than late learners, and for native learners,

accuracy was similar for control and violation sentences, $F(1, 36) = 0.3$, $p = .59$. However, late learners correctly identified control sentences more reliably than sentences containing semantic violations, $F(1, 36) = 6.5$, $p = .015$. It is important to emphasize, however, that in spite of overall lower proficiency the late learners were quite accurate in the sentence judgment task and well above chance. ERP data were only analyzed for sentences that participants responded to correctly.

ERP Data

Visual inspection of the ERP waveforms across the various conditions, seen in Figure 2, suggested that both native and late learners showed enhanced N400 responses to semantic violations relative to control sentences. However, the N400 appeared to have an earlier onset and greater amplitude for native than for late learners. Subsequent to the N400, additional violation effects were observed in both groups from approximately 700–900 msec, consisting of a LAN and a more posterior positivity. The scalp distributions of these effects (seen in Figure 3) seemed to differ between groups, however, with native learners showing the negativity more prominently while late learners' topographies were dominated by the posterior-distributed positivity.

N400 Amplitude—Figure 4 shows the difference waves for the native and late learner groups. Both groups showed an enhanced negativity peaking around 450 msec, as is typical of the N400 effect. The scalp distribution of this effect, as seen in Figure 3, was maximal over midline electrodes for both groups. To analyze the amplitude and scalp distribution of the N400, we performed LME modeling on mean amplitudes over a 100-msec time window centered on the peak of the N400 violation effect, from 400 to 500 msec. The optimal LME model included factors Condition, Proficiency, Group, and ROI, as well as all two- and three-way interactions between these factors, and the four-way interaction $\text{Proficiency}_{\text{residualized}} \times \text{Condition} \times \text{ROI} \times \text{Group}$, as well as by-subject random intercepts and slopes for ROI. This model is shown in Table 1A.

The mean ERP amplitude in the violation condition was more negative than in the control condition ($M_{\text{Violation}} = -1.36$ μV, $M_{\text{Control}} = 0.17$ μV, difference = −1.53 μV). The Condition × ROI interaction reflected the fact that the violation–control difference was largest over the vertex of the scalp, as is typical of the N400. To confirm this observation, we conducted a post hoc analysis contrasting the three levels of the anteriority gradient as well as the three levels of the laterality gradient. Results are provided in Table 1B. Examination of the means for each ROI suggested that the condition effect at central and posterior regions of the midline and right hemisphere were similar. We thus made five additional, more specific comparisons (see Table 1B). The post hoc analysis revealed greater N400 amplitudes at midline central and posterior ROIs, which is consistent with the classic N400 associated with semantic violations.

The significant Condition × ROI × Group interaction, which survived vector scaling,[4] indicates that the magnitude of the condition-related negativity differed between native and late learners at certain scalp sites. We conducted post hoc comparisons on these "difference-of-difference" scores (calculated as $\text{L2}_{(\text{Violation–Control})} - \text{L1}_{(\text{Violation–Control})}$) among levels.

The results, shown in Table 1C, indicated that the condition effect was greater for native speakers than late learners at anterior and central scalp sites, as is apparent in Figure 3.

The four-way interaction $\text{Proficiency}_{\text{residualized}}$ × Condition × ROI × Group remained highly significant after vector scaling. This indicated that proficiency affected N400 amplitudes differently in each group. To explore this four-way interaction, we inspected the native learner and late learner data separately. In each case, we fitted a model with a three-way interaction Proficiency × Condition × ROI (proficiency was mean-centered but not residualized here, because analyses were restricted to single groups) and by-subject random intercepts and slopes for ROI.

Results for the native learner analysis are shown in Table 2A. The three-way interaction is graphed in Figure 5A. Each panel graphs the Proficiency × Condition interaction for each ROI, as well as the proficiency simple effects for the control and violation conditions. Across all ROIs, the observable trend was for a relatively flat line relating violation word amplitude to proficiency, but increasingly negative amplitudes for control words for lower-proficiency participants. This led to larger N400 violation–control differences in higher-proficiency participants. To narrow the locus of the proficiency effect in native learners, we conducted a post hoc analysis by comparing, between levels of the anteriority gradient and between levels of the laterality gradient, the slope of the line predicting the (violation–control) difference as a function of proficiency. Results are shown in Table 2B. In brief, the magnitude of the proficiency effect was greater at left regions across the levels of anteriority.

For late learners, the three-way interaction Proficiency × Condition × ROI did not reach significance, nor did the Proficiency × ROI interaction. The Condition × ROI interaction was, however, significant. The Proficiency × Condition interaction was also significant. The optimal model is summarized in Table 2C. The results of a post hoc analysis of the Condition × ROI interaction provided in Table 2D indicate that the N400 was maximal at centro-posterior midline regions. Figure 5B depicts the Proficiency × Condition interaction, which reflects the amplitude of the N400 violation effect becoming more negative as proficiency increases. Similarly to the case with native learners, this effect appears to have been driven primarily not by the amplitude of the violation words (which was relatively flat across the proficiency range) but rather by more negative amplitudes for control words in less proficient late learners.

The foregoing analyses indicated that proficiency had an effect on N400 mean amplitudes independently of L1/ L2 status, because the variance in proficiency associated with group membership had been removed through residualization. In this model, however, group was still confounded with proficiency; the observed effects attributed to group (i.e., Condition × ROI × Group and ROI × Group interactions) could just as well be attributed to that part of

[4]Because the additive nature of ANOVAs is incompatible with the multiplicative nature of interactions, it is possible that the interactions involving ROI are spurious. McCarthy and Wood (1985; see also Dien & Santuzzi, 2005) developed a scaling method that addresses this potential problem called vector scaling. For each participant and each condition, mean amplitudes are scaled by the square root of the sum of the squared mean amplitudes, i.e., $X_{ij}/\sqrt{\Sigma}(X_{ij}^2)$, where X_{ij} is the amplitude for subject i in condition j. If the scalp region by condition interaction remains significant after rescaling then one may be more confident in the veracity of the effect, under certain conditions (Urbach & Kutas, 2002, 2006).

proficiency that correlated with group. As a consequence, we cannot draw a strong conclusion from these analyses as to whether native/late learner status affects N400 mean amplitudes independently of proficiency. We thus conducted a further analysis in which the effects of proficiency were first removed from the ERP data, and then the effect of group was assessed. Specifically, we fitted a linear model with the three-way interaction Proficiency × Condition × ROI (mean-centered but, critically, unresidualized), took the model residuals, and then fitted an LME model on these with the three-way interaction Condition × ROI × Group as well as by-subject random intercepts and slopes for ROI. This procedure should have the effect of first removing all variance associated with proficiency from the data before assessing the influence of group. Results are provided in Table 3. The main effect of condition and the interactions of Condition × Group and Condition × ROI × Group were not significant, suggesting that group did not significantly affect N400 mean amplitudes independently of proficiency (i.e., once proficiency was accounted for). The trend, however, remained similar: the negativity was greater for native learners than late learners at anterior and central scalp sites.

Finally, given that age and years of education differed between groups, our results may have been confounded by these variables.[5] To test this, we first removed the effects of age and years of education on N400 mean amplitudes by fitting a linear model predicting N400 mean amplitudes on the basis of age and years of education. We then refitted our mixed effects model on the residuals of this. The pattern of main effects and interactions and their significance remained the same as in our initial model, indicating that the effects were not because of differences in age or education between the groups.

To summarize, violation–control N400 amplitudes were overall larger in L1 than L2 learners. In both groups, larger N400 amplitudes were found in higher-proficiency participants, and once the effects of proficiency were accounted for, L1/L2 learner status did not predict N400 differences. However, differences in the effects of proficiency were seen between groups. Specifically, the effects of proficiency were predominant over the left hemisphere in L1 learners but widely distributed over the scalp in L2 learners.

N400 Onset Latency—In addition to the amplitude differences, examination of the difference waveforms shown in Figure 4 suggested that the onset of the N400 violation effect was earlier for native learners. To test this, we modeled the 20% fractional area latency of the N400 violation effect (difference waveform, calculated as violation–control) in the 200–600 msec time window. This was calculated as the time at which 20% of the total mean amplitude was obtained within this time window, which was chosen to capture the observed epoch over which the violation and control waveforms differed (Hansen & Hillyard, 1980). We fitted an LME model predicting fractional area latency, on the basis of fixed effects proficiency (residualized), group, and ROI and by-subject random intercepts and slopes for ROI. The optimal model contained only main effects of group and ROI.

[5]Age × Years of Education: $r = .49$, $\beta = 1.2$ (as age increases, so does years of education), $t(34) = 3.3$, $p = .003$. Age × Group: $r = .59$, $\beta = 9.9$ (L2s are that much older than L1s), $t(34) = 4.3$, $p = .0001$. Years of Education × Group: $r = .46$, $\beta = 3.2$ (L2s have that much more years of education than L1s), $t(34) = 3.1$, $p = .004$. Age × Residualized Proficiency: $r = .13$, $\beta = -0.3$, $t(34) = -0.8$, $p = .46$. Years of Education × Residualized Proficiency: $r = .1$, $\beta = 0.59$, $t(34) = 0.6$, $p = .55$. Group × Residualized Proficiency: $r = .04$, $\beta = 1.6$, $t(34) = 0.2$, $p = .81$.

A main effect of group confirmed that the onset of the N400 violation effect was earlier in natives than in late learners, M_{L1} = 311 msec, M_{L2} = 339 msec, difference = 28 msec, $F(1, 2287) = 6.3$, $p = .03$. There was also a main effect of ROI, $F(8, 2287) = 2.1$, $p = .01$. A post hoc analysis comparing the three levels of the anteriority and the laterality gradients revealed that the only reliable N400 onset difference was between anterior and central ROIs, $M_{anterior}$ = 324 msec, $M_{central}$ = 340 msec, $t_{(2287)} = 2.7$, $p = .04$, two-tailed and Bonferroni's corrected for six comparisons.

As in the previous analyses, to assess the relative contributions of group and proficiency, we fitted a linear model on mean fractional area latencies as a function of Proficiency × ROI (mean-centered, but not residualized) and then fitted an LME model on the residuals with ROI and group as fixed effects and by-subject random intercepts and slopes for ROI. As Table 4A shows, there were no significant interactions or main effects. We performed a similar analysis to determine the potential contribution of proficiency (not residualized) to the effect, once the variance associated with group was regressed out. Results are provided in Table 4B. Similarly to group, proficiency did not independently affect N400 latencies. Although group appears to have accounted for a greater portion of the overall variability (MSS_{Group} = 1598.5 and $MSS_{ROI \times Group}$ = 486.1 vs. $MSS_{Proficiency}$ = 164.1 and $MSS_{ROI \times Proficiency}$ = 355.3), these results suggest that the effect of these two variables on N400 onset latencies cannot be unequivocally dissociated.

Finally, as in the N400 amplitude analysis, we checked whether the group effect was confounded with age and years of education. After accounting for these potential confounds, the effect of group became marginally significant, $F(1, 2287) = 3.2$, $p = .07$. Thus, although the N400 onset appears faster in native speakers than late learners, this trend may not be replicable.

In summary, N400 onset latencies were earlier in L1 than L2 learners. However, onset latencies were not significantly affected by proficiency, suggesting that the timing of N400 onset is more closely tied to L1/L2 learner status.

Late Effects of Condition—Subsequent to the N400, the scalp topography of the difference waves appeared to differ between groups. As seen in Figures 2 and 3, the dominant feature of the scalp maps of native speakers was a late negativity from approximately 600–900 msec, maximal over left and midline anterior sites, whereas for late learners a late positivity beginning around 650 msec and largest over posterior right and mid-line sites was most salient. However, the posterior positivity and anterior negativity were visible, albeit diminished, in the native and late learner groups, respectively. To explore these effects, we performed LME modeling on mean ERP amplitudes from 700 to 900 msec. The optimal model is shown in Table 5A; the results were the same after controlling for age and years of education as described for the previous analyses.

The main effects of Condition and Group were not significant. The effect of ROI, which was reliable, was modulated by condition type. The difference between the violation and control conditions at anterior scalp regions appeared more negative than at central and posterior sites and the right posterior region was the most positive on the scalp.

Post hoc analyses, shown in Table 5B, compared the three levels of anteriority and laterality gradients. The results confirmed the strongest negativity at anterior sites (all *p* values smaller than .0001). From left to right, a similar, although shallower, gradient also seemed to be present. Although differences between L and R scalp sites were reliable, neither the differences between the L and M, nor the R and M, regions approached statistical significance.

There was also a significant Condition × Group interaction. Reflecting the overall more prevalent negative potentials across the scalp in native learners and more widespread positive potentials in late learners, across ROIs the violation–control difference was more negative for native speakers, $M_{L1} = -0.9$ μV, and more positive for late learners, $M_{L2} = 0.2$ μV.

Finally, to assess whether group affected the late positivity independently of proficiency, we repeated the procedure used for previous measures and fitted a linear model to the mean amplitudes with ROI, condition, and proficiency, then, on the model residuals, performed an LME regression with the three-way interaction Proficiency × Condition × ROI with by-subject random intercepts and slopes for ROI. The only significant effect that emerged from this model was a Condition × Group interaction [$F(1, 4651) = 50.0$, $p < .0001$] where the violation minus control difference was more negative for native speakers than late learners ($M_{L1(Violation– Control)} = -0.3$ μV, $M_{L2(Violation–Control)} = 0.3$ μV, difference of differences = 0.6 μV). These results are similar to the ones obtained from the model where group and proficiency were confounded, although the difference is smaller when the effect of proficiency is first removed.

To summarize, from 700 to 900 msec, both L1 and L2 learners showed an anterior negativity and a posterior positivity. The anterior negativity was more prominent across the scalp in L1 learners, whereas the posterior positivity was more widespread in L2 learners. These late effects of semantic violations were not affected by proficiency however. Rather, they reflect differences in the way first versus later-learned languages are processed.

DISCUSSION

The aims of this study were to (1) examine the effects of language proficiency on the ERP components elicited by lexical semantic violations during sentence processing, in both native and late learners, and (2) to statistically separate the effects of proficiency on ERPs from the effects of native versus late language acquisition. If proficiency was a critical factor in determining brain activation, then significant and similar effects of proficiency would be obtained across both groups. On the other hand, if late acquisition leads to a fundamentally different brain organization for language, then group differences in the N400 would hold even when proficiency was considered in the analysis.

N400 Amplitude

An N400 was reliably elicited by semantic violations in both groups. The timing and scalp distribution of this effect in native speakers was comparable to that obtained using the same stimuli in a previous study (Newman et al., 2007). The amplitude of the N400 violation

effect was significantly greater for native speakers than late learners of English at anterior and central scalp regions, replicating previous findings (Hahne et al., 2006; Moreno & Kutas, 2005; Ojima et al., 2005; Weber-Fox & Neville, 1996). Proficiency affected N400 amplitudes independently of group, with the size of the N400 violation effect increasing as proficiency increased. Although previous studies had shown greater N400s in more proficient and/or earlier learning L2 learners (Rossi et al., 2006; Moreno & Kutas, 2005; Ojima et al., 2005; Phillips et al., 2004), this study gives us greater insight into the nature of the proficiency effects. Because we treated proficiency as a measure that varied continuously across both L1 and L2 learners and residualized it with respect to group, we can conclude that proficiency affects N400 amplitude similarly in L1 and L2 learners.

The origin of these N400 differences was largely because of the amplitude of the response to semantically congruous words. Lower-proficiency English speakers, regardless of whether they were native or late learners, showed more negative potentials than those with higher proficiency; the negativities elicited by incongruous words were of similar amplitude in both groups. Thus, the attenuated N400 violation effects observed for late learners in this and in other studies appear to stem from increased costs of semantic integration for open class words generally in lower-proficiency language users, rather than from differences in how violations are processed. Different interpretations of the N400 have been put forward (Lau et al., 2008; Kutas & Federmeier, 2000); the greater integration cost in lower-proficiency speakers may be because of less efficient lexical access and/or poorer ability to predict words in well-formed sentences.

Although both native and late learners showed a similar relationship between proficiency and N400 amplitude, group differences were found in the scalp distribution of this relationship. In native speakers, the effect was largest over left scalp sites, although observable at virtually every ROI, whereas in late learners it was widely distributed across the scalp. The N400 is known to have a distributed set of neural generators (Lau et al., 2008), and we had insufficient data to expect reliable source localization here. However, the evidence is consistent with differential distribution of proficiency-related brain activity in native compared with late learners.

N400 Latency

The estimated onset of the N400 was also earlier for native than late learners but was not clearly related to English proficiency in either group—when group and residualized proficiency were both included in the LME model, group was a significant predictor but proficiency was not. At the same time, the results of the analyses in which group was first regressed out and then the residuals fitted against proficiency, and vice versa, suggested that the effects of the two variables could not be unequivocally dissociated. Furthermore, latency appeared to be marginally influenced by age and years of education. Thus generally, our findings are consistent with previous work showing increased latency of ERP components in L2 learners (Rossi et al., 2006; Moreno & Kutas, 2005; Ojima et al., 2005; Phillips et al., 2004). However, in this previous work, higher proficiency was shown to be associated with earlier N400 peak latency, even once group was accounted for (Moreno & Kutas, 2005). There are at least two possible explanations why we found only evidence for the influence of

group on N400 latency in this study. Onset and not peak latency was used, as the N400 difference waves did not have a single clear peak in the present data (possibly because of overlapping word-offset ERP components). N400 onset latency may be less strongly related to proficiency than peak latency. Second, the bilinguals who showed effects of proficiency learned their L2 at a comparatively early age (mean ages of 8 and 12, respectively). The late learners in our study did not really learn English until immersed in it after the age of 18. The effects of proficiency on N400 latency may be restricted to earlier learners.

Late Effects of Semantic Violations

Both groups showed an anterior negativity along with a posterior positivity in the 700–900 msec time range. However, the anterior negativity was more prominent for the native learners, whereas the posterior positivity was more prominent in late learners. These differences in scalp distribution were not influenced by English proficiency, when L1/L2 learner status was accounted for. Late anterior negativities have been associated with working memory in sentence processing (Vos, Gunter, Kolk, & Mulder, 2001; Münte, Schiltz, & Kutas, 1998). The present results may suggest that native learners rely more on this on-line storage, perhaps to recheck the preceding input, whereas late learners are more strongly disrupted by discrepancies between input and expectations.

Late positivities are often reported in response to semantic anomalies (van de Meerendonk, Kolk, Chwilla, & Vissers, 2009; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Moreno & Kutas, 2005; Ojima et al., 2005; Coulson & Van Petten, 2002; Juottonen & Revonsuo, 1996). Retrospectively, we note that, in our previous study using these same stimuli (Newman et al., 2007), a small posterior positivity was present for semantic violations, although statistical analyses were not performed on that time window. Kolk and colleagues (van de Meerendonk et al., 2009) have suggested that the P600 may reflect a general purpose monitoring process that compares predictions with actual input across a variety of domains, including syntax, semantics, and other sequences including mathematics and music. They further propose that strong (unresolvable) expectancy violations elicit a P600, whereas weaker (resolvable) violations of expectancy elicit an N400. In this study, the larger positivity for late learners may indicate that they formed relatively strong expectations of sentence continuations and are more challenged when input is highly inconsistent with their expectations. Late learners may have less “flexibility” in considering alternate senses of a word or interpretations of a sentence (e.g., a metaphorical interpretation). However, previous studies of earlier-learning bilinguals found larger late positivities elicited by semantic violations in native learners (Ojima et al., 2005) and in bilinguals’ dominant language (Moreno & Kutas, 2005). The reasons for this discrepancy may again relate to the AoA of L2.

Two caveats are important here. One is that, although there were group differences in the relative prominence of the anterior negativity and posterior positivity, both potentials were observable in both groups. This may suggest differential reliance on the two processes indexed by these potentials but not qualitatively distinct processing between native and late learners. Secondly, an inherent limitation of scalp topography data is that only one of these two processes may actually differ between groups, for example, a larger anterior negativity

in native learners would tend to reduce the prominence of the posterior positivity in this group, even if the strength of the generators underlying the positivity was similar across both groups.

Methodological Considerations

Decorrelating Proficiency and Group Status—As expected, native speakers had significantly higher proficiency than late learners. Because of the collinearity of proficiency and group, assessing the independent influence of each variable was impossible. One approach to this would be stepwise regression (e.g., Moreno & Kutas, 2005). However, although main effects can be examined in this way, interpretation of interactions is difficult at best, and in the present data the primary questions were centered on interactions between condition, ROI, proficiency, and group. Our first approach to this problem was to use the residual variance in proficiency that was not accounted for by group membership, making the implicit assumption that any overall difference between groups is because of L1/L2 learner status. This may or may not be correct, because in sampling, limited groups of individuals there may be an overall difference in proficiency unrelated to learner status. Thus, we also took a second approach, in which we first removed the effects of proficiency from the ERP data and then determined whether any residual variance from this model (i.e., not predicted by proficiency) was predicted by group. This, in effect, asks whether, once proficiency was accounted for, L1/L2 learner status added any predictive value. In this case, the answer was “no”—neither the main effect of group nor any interactions with this variable were significant. Thus both analyses found similar effects of proficiency on N400 amplitudes within each group.

Measuring Language Proficiency—We chose to use a standardized test of English ability, the TOAL-3, that was developed on a normative sample of native English learners aged up to 24 years, and is sensitive to a range of abilities among this group. This is in contrast to clinical tests used in some previous studies on which most native speakers score at ceiling. Nevertheless, it is important to consider that the choice of a proficiency measure may affect the outcome of a study. It will also be important in future work to consider the question of what the construct of “proficiency” really is—likely a proxy for the amount, quality, and type of language input—and whether the factors that modulate proficiency are similar or different in L1 and L2. These variables can be difficult to accurately assess retrospectively, although training studies with artificial, miniature, and real languages provide opportunities to more precisely measure and control these factors. Language dominance also likely plays a role (Moreno & Kutas, 2005), as one may become more proficient in the language that one uses most heavily, even if it is not an L1. Future work will be needed to explore the relationship between dominance and proficiency.

Conclusions

Our results suggest that different indices of lexical semantic processing are differently affected by language proficiency and by late versus native acquisition. The amplitude of the N400 to semantically congruent words was larger in lower-proficiency English speakers regardless of native/late learner status. This indicates that lexical semantic integration during sentence processing is affected by fluency, not but L2 learning specifically. Previous

observations of lower N400 amplitudes in L2 learners are likely attributable to lower average proficiency.

In contrast, the delayed onset of the N400 violation effect and the relative balance of anterior negativity versus posterior positivity subsequent to the N400 were not clearly attributable to differences in proficiency. This difference may thus be attributable to slower lexical semantic processing in L2 learners, independent of proficiency obtained. Further, L1 learners may rely more on working memory when attempting to resolve lexical semantic violations, reflected by anterior negativities after the N400, whereas L2 learners' processing is more disrupted as evidenced by a more predominant late positivity. Again, these differences appear to reflect fundamental L1/L2 processing differences as they were unaffected by proficiency.

Acknowledgments

This study was supported by NIH NIDCD DC00128 to H. J. N.; NSF SBR-9905273, NIH R01 MH58189, NIH R01 HD049347, and Army DAMD-17-93-V-3018/3019/3020 and DAMD-17-99-2-9007 to M. T. U.; A. J. N. was supported by an NSERC Discovery Grant and the Canada Research Chairs program. We are grateful to Diane Waligura, Matthew Moffa, Claudia Brovetto, Linda Heidenreich, Kara Morgan-Short, Kaori Ozawa, Roumayana Pancheva, Jackie Schachter, Karsten Steinhauer, Ray Vukevich, Jill Weisberg, and Harriet Wood-Bowden for their assistance with this study.

References

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59:390–412.

Bagiella E, Sloan RP, Heitjan DF. Mixed-effects models in psychophysiology. Psychophysiology. 2000; 37:13–20. [PubMed: 10705763]

Bates DM. Fitting linear mixed models in R. R News. 2005; 5:27–30.

Bates, DM.; Maechler, M.; Bolker, B. R package version 0.999375-39. The Comprehensive R Archive Network (CRAN): The Institute of Statistics and Mathematics of the Wirtshaftsuniversität Wien (WU); 2009. lme4: Linear mixed-effects models using S4 classes.

Birdsong D, Molis M. On the evidence for maturational constraints in second-language acquisition. Journal of Memory and Language. 2001; 44:235–249.

Cohen J. The cost of dichotomization. Applied Psychological Measurement. 1983; 7:249–253.

Coulson S, Van Petten C. Conceptual integration and metaphor: An event-related potential study. Memory & Cognition. 2002; 30:958–968. [PubMed: 12450098]

Davidson DJ, Indefrey P. An inverse relation between event-related and time-frequency violation responses in sentence processing. Brain Research. 2007; 1158:81–92. [PubMed: 17560965]

Dien, J.; Santuzzi, AM. Application of repeated measures ANOVA to high-density ERP datasets: A review and tutorial. In: Handy, TC., editor. Event-related potentials. A methods handbook. Cambridge, MA: MIT Press; 2005. p. 57-82.

Elston-Güttler KE, Paulmann S, Kotz SA. Who's in control? Proficiency and L1 influence on L2 processing. Journal of Cognitive Neuroscience. 2005; 17:1593–1610. [PubMed: 16269099]

Flege J, Yeni-Komshian GH, Liu S. Age constraints on second-language acquisition. Journal of Memory and Language. 1999; 41:78–104.

Friederici AD, Steinhauer K, Pfeifer E. Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. Proceedings of the National Academy of Sciences, USA. 2002; 99:529–534.

Gelman, A.; Hill, J. Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press; 2007.

Hahne A, Mueller JL, Clahsen H. Morphological processing in a second language: Behavioral and event-related brain potential evidence for storage and decomposition. Journal of Cognitive Neuroscience. 2006; 18:121–134. [PubMed: 16417688]

Hammill, DD.; Brown, VL.; Larsen, SC.; Wiederholt, JL. Test of Adolescent and Adult Language, third edition (TOAL-3). 3. Austin, TX: Pro-Ed; 1994.

Hansen JC, Hillyard SA. Endogenous brain potentials associated with selective auditory attention. Electroencephalography and Clinical Neurophysiology. 1980; 49:277–290. [PubMed: 6158404]

Johnson JS, Newport EL. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language1. Cognitive Psychology. 1989; 21:60–99. [PubMed: 2920538]

Juottonen K, Revonsuo A. Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. Cognitive Brain Research. 1996; 4:99–107. [PubMed: 8883923]

Kotz SA, Elston-Güttler KE. The role of proficiency on processing categorical and associative information in the L2 as revealed by reaction times and event-related brain potentials. Journal of Neurolinguistics. 2004; 17:215–235.

Kuperberg GR, Kreher DA, Sitnikova T, Caplan DN, Holcomb PJ. The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. Brain and Language. 2007; 100:223–237. [PubMed: 16546247]

Kutas M, Federmeier KD. Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences. 2000; 4:463–470. [PubMed: 11115760]

Lau EF, Phillips C, Poeppel D. A cortical network for semantics: (De)constructing the N400. Nature Reviews Neuroscience. 2008; 9:920–933.

Lenneberg, E. Biological foundations of language. New York: Wiley; 1967.

McCarthy G, Wood CC. Scalp distribution of event-related potentials: An ambiguity associated with analysis of variance models. Electroencephalography and Clinical Neurophysiology. 1985; 62:203–208. [PubMed: 2581760]

Meschyan G, Hernandez AE. Impact of language proficiency and orthographic transparency on bilingual word reading: An fMRI investigation. Neuroimage. 2006; 29:1135–1140. [PubMed: 16242351]

Midgley KJ, Holcomb PJ, Grainger J. Language effects in second language learners and proficient bilinguals investigated with event-related. Journal of Neurolinguistics. 2009; 22:281–300. [PubMed: 19430590]

Moratti S, Clementz BA, Gao Y, Ortiz T, Keil A. Neural mechanisms of evoked oscillations: Stability and interaction with transient events. Human Brain Mapping. 2007; 28:1318–1333. [PubMed: 17274017]

Moreno EM, Kutas M. Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish–English bilinguals. Brain Research, Cognitive Brain Research. 2005; 22:205–220. [PubMed: 15653294]

Morgan-Short K, Sanz C, Ullman MT. Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. Language Learning. 2010; 60:154–193. [PubMed: 21359123]

Mueller JL, Hahne A, Fujii Y, Friederici AD. Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. Journal of Cognitive Neuroscience. 2005; 17:1229–1244. [PubMed: 16197680]

Mueller JL, Oberecker R, Friederici AD. Syntactic learning by mere exposure: An ERP study in adult learners. BMC Neuroscience. 2009; 9:1–10.

Münte TF, Schiltz K, Kutas M. When temporal terms belie conceptual order. Nature. 1998; 395:71–73. [PubMed: 9738499]

Newman AJ, Ullman MT, Pancheva R, Waligura DL, Neville HJ. An ERP study of regular and irregular English past tense inflection. Neuroimage. 2007; 34:435–445. [PubMed: 17070703]

Newman-Norlund RD, Frey SH, Petitto LA, Grafton ST. Anatomical substrates of visual and auditory miniature second-language learning. Journal of Cognitive Neuroscience. 2006; 18:1984–1997. [PubMed: 17129186]

Ojima S, Nakata H, Kakigi R. An ERP study of second language learning after childhood: Effects of proficiency. Journal of Cognitive Neuroscience. 2005; 17:1212–1228. [PubMed: 16197679]

Pakulak E, Neville HJ. Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. Journal of Cognitive Neuroscience. 2010; 22:2728–2744. [PubMed: 19925188]

Perani D, Abutalebi J, Paulesu E, Brambati S, Scifo P, Cappa SF, et al. The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. Human Brain Mapping. 2003; 19:170–182. [PubMed: 12811733]

Phillips NA, Segalowitz N, Brien IO, Yamasaki N. Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. Journal of Neurolinguistics. 2004; 17:237–262.

Pinheiro, JC.; Bates, DM. Mixed-effects models in S and S-PLUS. New York: Springer; 2000.

Pritchett S, Zilberg E, Xu ZM, Myles P, Brown I, Burton D. Peak and averaged bicoherence for different EEG patterns during general anaesthesia. Biomedical Engineering Online. 2010; 9:76. [PubMed: 21092128]

Rossi S, Gugler MF, Friederici AD, Hahne A. The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. Journal of Cognitive Neuroscience. 2006; 18:2030–2048. [PubMed: 17129189]

Tremblay, A. LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. R package version 1.6.3. 2011. http://cran.r-project.org/web/packages/LMERConvenienceFunctions/index.html

Tremblay A, Tucker BV. The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. The Mental Lexicon. 2011; 6:302–324.

Urbach TP, Kutas M. The Intractibility of scaling scalp distributions to infer neuroelectric sources. Psychophysiology. 2002; 39:791–808. [PubMed: 12462507]

Urbach TP, Kutas M. Interpreting event-related brain potentials (ERP) distributions: Implications of baseline potentials and variability with application to amplitude normalization by vector scaling. Biological Psychology. 2006; 72:333–343. [PubMed: 16446023]

van de Meerendonk N, Kolk HHJ, Chwilla DJ, Vissers CTWM. Monitoring in language perception. Language and Linguistics Compass. 2009; 3:1211–1224.

Vos SH, Gunter TC, Kolk HH, Mulder G. Working memory constraints on syntactic processing: An electrophysiological investigation. Psychophysiology. 2001; 38:41–63. [PubMed: 11321620]

Wartenburger I, Heekeren HR, Abutalebi J, Cappa SF, Villringer A, Perani D, et al. Early setting of grammatical processing in the bilingual brain. Neuron. 2003; 37:159–170. [PubMed: 12526781]

Weber-Fox CM, Davis LJ, Cuadrado E. Event-related brain potential markers of high-language proficiency in adults. Brain and Language. 2003; 85:231–244. [PubMed: 12735941]

Weber-Fox CM, Neville HJ. Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. Journal of Cognitive Neuroscience. 1996; 8:231–256. [PubMed: 23968150]

Wierda SM, van Rijn H, Taatgen NA, Martens S. Distracting the mind improves performance: An ERP study. PloS One. 2010; 5:e15024. [PubMed: 21124833]

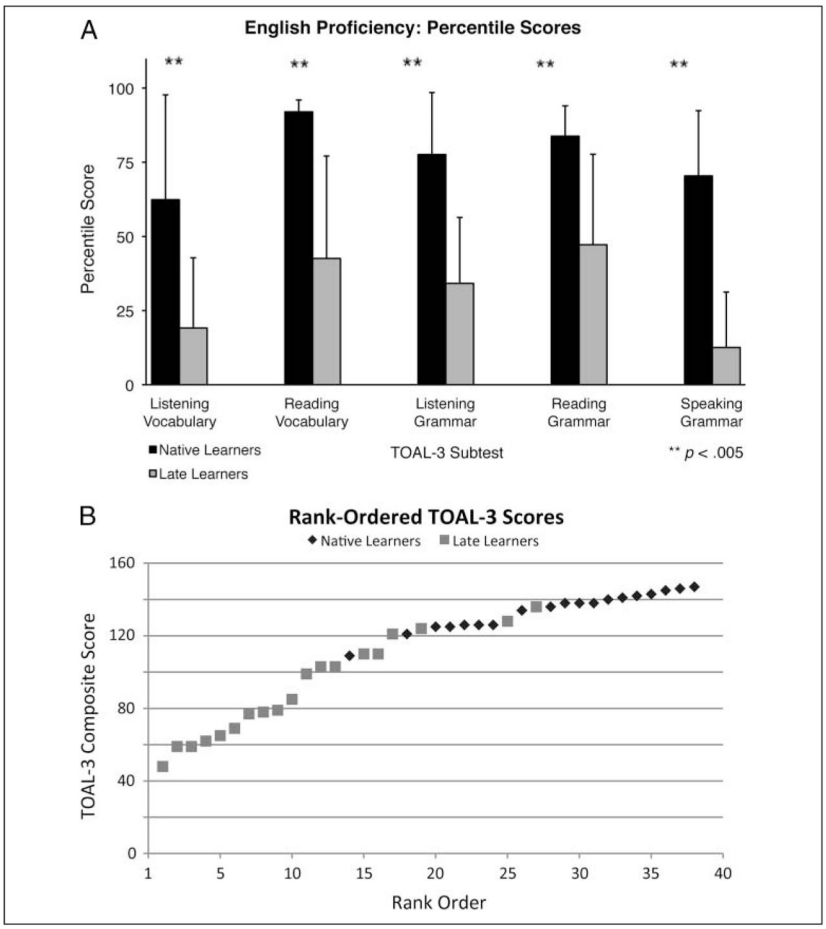

Figure 1.

(A) TOAL-3 subtest scores for native and late learners, expressed as percentiles based on published norms. Error bars represent standard deviations. (B) Rank-ordered TOAL-3 composite scores. Maximum possible composite TOAL-3 score was 155.

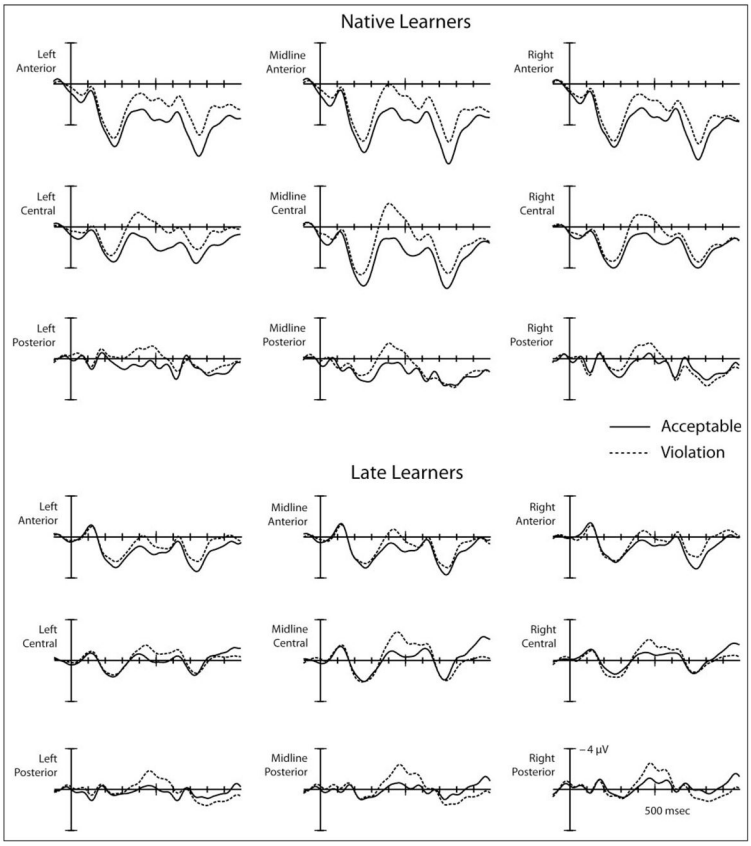

Figure 2.

ERP waveforms averaged across groups of electrodes, for lexical semantic violations and control words. Negative voltage is plotted upward.

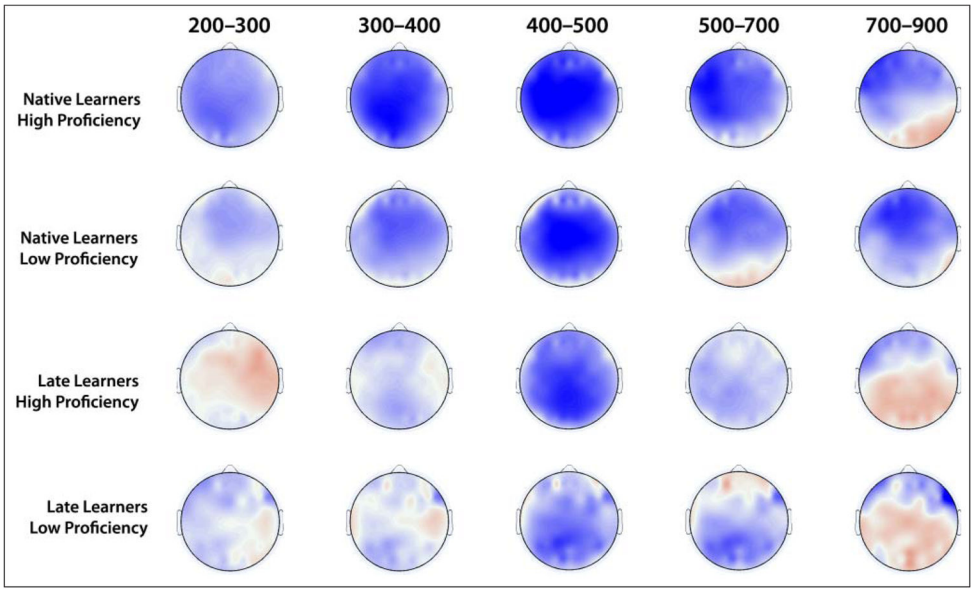

Figure 3.

Scalp voltage maps showing the difference between violation and control words, averaged over selected time windows. Scale is −3 μV (blue) to +3 μV (red). Participants were split into “high”- and “low”-proficiency subgroups based on a median split of composite TOAL-3 scores within each group. Although in the statistical analyses proficiency was treated as a continuous variable rather than dichotomized, here the dichotomization serves to highlight differences that are consistent within each learner group, across variation in proficiency.

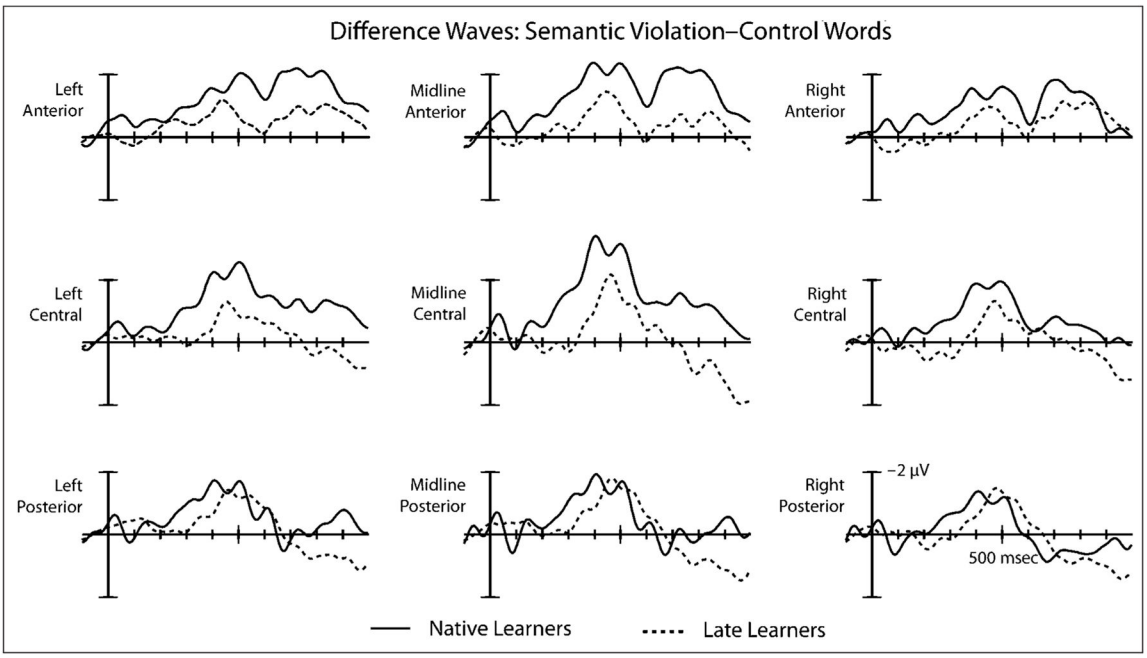

Figure 4.
ERP difference waveforms, computed as violation–control. Negative is plotted upward.

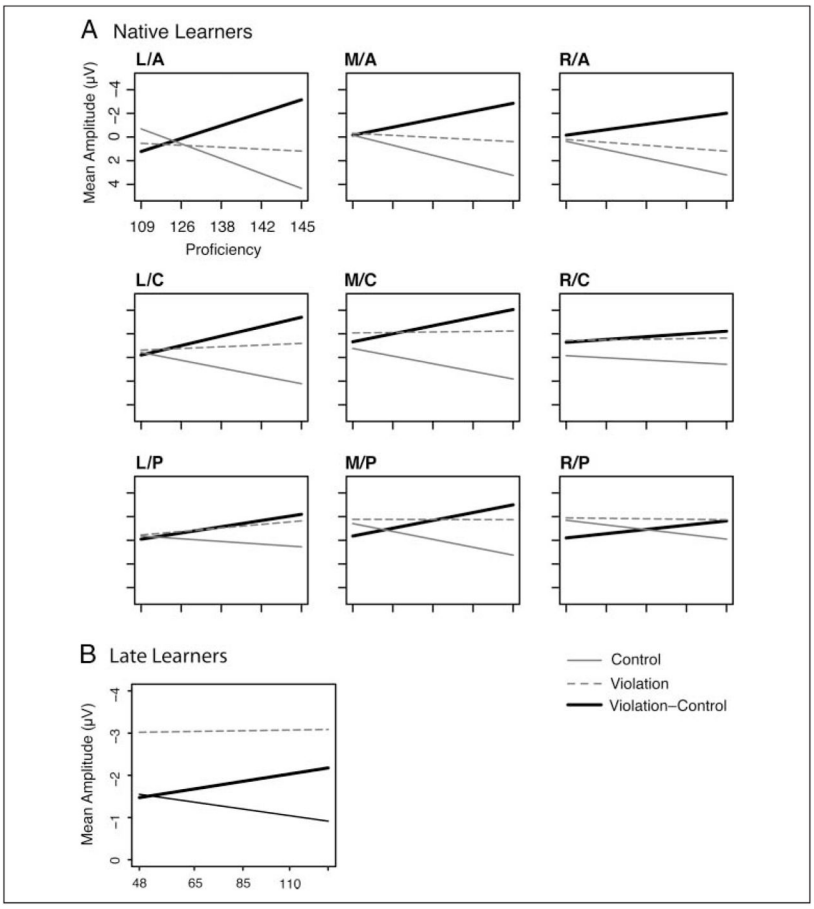

Figure 5.

(A) Proficiency × Condition interaction for each ROI as well as the proficiency simple effects for native speakers of English. The *x* axis is proficiency, and the *y* axis is amplitude (negative is plotted up). The solid black line is the two-way interaction, the solid gray line is the proficiency simple effect in the control condition, and the broken gray line is the simple effect of proficiency in the violation condition. (B) The Proficiency × Condition interaction for late learners of English. The *x*–*y* axes are proficiency and mean amplitude respectively. The solid and broken gray lines are proficiency for the control and violation conditions, respectively, and the solid black line is the violation minus control difference.

J Cogn Neurosci. Author manuscript; available in PMC 2015 May 28.

Table 1

ANOVA Table for N400 Amplitude in the 400–500 msec Time Window (Denominator Lower-bound *df* = 4299; Upper-bound *df* = 4703) and Post hoc Probability Values

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
A. ANOVA						
$Proficiency_{residualized}$	1	0.8	0.8	0.8	.37	.37
Condition	1	32.2	32.2	32.2	<.0001	<.0001
ROI	8	244.7	30.6	31.4	<.0001	<.0001
Group	1	0.1	0.1	0.1	.80	.80
$Proficiency_{residualized}$ × Condition	1	0.9	0.9	0.9	.33	.33
$Proficiency_{residualized}$ × ROI	8	9.2	1.2	1.2	.31	.31
Condition × ROI	8	188.8	23.6	24.2	<.0001	<.0001
$Proficiency_{residualized}$ × Group	1	3.3	3.3	3.4	.07	.07
Condition × Group	1	1.3	1.3	1.4	.24	.24
ROI × Group	8	41.0	5.1	5.3	<.0001	<.0001
$Proficiency_{residualized}$ × Condition × ROI	8	9.7	1.2	1.2	.27	.27
$Proficiency_{residualized}$ × Condition × Group	1	1.7	1.7	1.8	.18	.18
$Proficiency_{residualized}$ × ROI × Group	8	12.5	1.6	1.6	.12	.12
Condition × ROI × Group	8	68.3	8.5	8.8	<.0001	<.0001
$Proficiency_{residualized}$ × Condition × ROI × Group	8	37.8	4.7	4.9	<.0001	<.0001

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
B. Post hoc, Condition × ROI			
C vs. A	4.8	.003	.003
C vs. P	3.6	.03	.03
A vs. P	−1.2	.99	.99
M vs. L	4.2	.01	.01
M vs. R	6.2	.0002	.0002
L vs. R	2.1	.56	.56
M/C vs. M/P	2.5	.20	.20
M/C vs. R/P	6.4	<.0001	<.0001

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
M/C vs. R/C	7.6	<.0001	<.0001
M/P vs. R/C	3.1	.05	.05
M/P vs. R/P	4.4	.002	.002
C. Post hoc, Condition × ROI × Group			
C vs. A	1.2	.99	.99
C vs. P	−5.3	<.0001	<.0001
A vs. P	−3.8	.006	.006
M vs. L	0.1	.99	.99
M vs. R	−1.0	.99	.99
L vs. R	−1.1	.99	.99
L/C vs. L/A	−1.4	.99	.99
L/C vs. M/C	0.3	.99	.99

Post hoc probability values were Bonferroni's corrected for 11 comparisons in (B) and 9 comparisons in (C); *t* tests were two-tailed. A = Anterior; C = central; P = posterior; L = left; M = midline; R = right.

Table 2

ANOVA Table (A and C) and Post hocs (B and D) for N400 Amplitude in the 400–500 msec Time Window, Conducted Separately for Each Group

A. Native Learners, ANOVA (Denominator Lower Bound df = 1945, Upper Bound df = 2349)

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
Proficiency	1	0.1	0.1	0.1	.82	.82
Condition	1	1880.7	1880.7	933.0	<.0001	<.0001
ROI	8	226.3	28.3	14.0	<.0001	<.0001
ROI × Condition	8	174.3	21.8	10.8	<.0001	<.0001
Proficiency × Condition	1	147.9	147.9	73.4	<.0001	<.0001
Proficiency × ROI	8	47.3	5.9	2.9	.003	.003
Proficiency × Condition × ROI	8	41.9	5.2	2.6	.008	.008

B. Native Learners, Post hoc, $Proficiency_{residualized}$ × Condition × ROI× Group

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
C vs. A	−2.1	.22	.22
C vs. P	−0.3	.99	.99
A vs. P	−2.2	.16	.16
M vs. L	−1.7	.55	.55
M vs. R	−1.6	.63	.63
L vs. R	−3.4	.004	.004

C. Late Learners, ANOVA (Denominator Lower Bound df = 1960; Upper Bound df = 2364)

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
Proficiency	1	886.3	886.3	505.0	<.0001	<.0001
Condition	8	322.2	40.3	22.9	<.0001	<.0001
ROI	1	0.6	0.6	0.4	.55	.55
ROI × Condition	8	79.7	10.0	5.7	<.0001	<.0001
Proficiency × Condition	1	24.5	24.5	13.9	.0002	.0002

D. Late Learners, Post hoc, Condition × ROI

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
C vs. A	−2.9	.02	.02

D. Late Learners, Post hoc, Condition × ROI

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
C vs. P	−1.0	.99	.99
A vs. P	−3.6	.002	.002
M vs. L	−2.8	.03	.03
M vs. R	3.1	.01	.01
L vs. R	0.3	.99	.99

Post hoc probability values were Bonferroni's corrected for six comparisons; *t* tests were two-tailed. A = Anterior; C = central; P = posterior; L = left; M = midline; R = right.

Table 3

N400 ANOVA Table on Residualized Mean Amplitudes (Denominator Lower Bound *df* = 4236; Upper Bound *df* = 4640)

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
Condition	1	5.6	5.6	3.5	.06	.06
ROI	8	1.0	0.13	0.1	.99	.99
Group	1	1.7	1.7	1.1	.3	.3
Condition × ROI	8	9.4	1.2	0.7	.66	.66
Condition × Group	1	5.6	5.6	3.5	.06	.06
ROI × Group	8	6.0	0.8	0.5	.86	.86
Condition × ROI × Group	8	21.3	2.7	1.7	.1	.1

Table 4

Contribution of Group (A) and Proficiency (B) Alone on N400 Onset Latencies

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
A. Group Alone (Denominator Lower bound df = 1876; Upper Bound df = 2280)						
ROI	8	3012.3	376.5	0.5	.82	.82
Group	1	1598.5	1598.5	2.3	.13	.13
ROI × Group	8	3888.8	486.1	0.7	.69	.69
B. Proficiency Alone (Denominator Lower Bound df = 1875; Upper Bound df = 2279)						
ROI	8	2538.0	317.3	0.5	.88	.88
Proficiency	1	164.1	164.1	0.2	.62	.62
ROI × Proficiency	8	2842.5	355.3	0.5	.84	.84

Note that proficiency was mean-centered but not residualized. 3.6% of the data were trimmed.

Table 5

ANOVA Table (A) and Post hoc Test Results (B) for Mean Amplitude in the 700–900 msec Time Window

A. ANOVA (Denominator Lower Bound df = 4244; Upper Bound df = 4648)

Coefficient	df	SumSq	MeanSq	F	p (Lower Bound)	p (Upper Bound)
Condition	1	3.6	3.6	2.3	.13	.13
ROI	8	110.9	13.9	9.0	<.0001	<.0001
Group	1	0.1	0.1	0.01	.92	.92
Condition × ROI	8	432.0	54.0	34.9	<.0001	<.0001
Condition × Group	1	7.4	7.4	4.8	.029	.029

B. Post hoc Analysis of the Condition× ROI Interaction

Comparisons between ROIs	t	p (Lower Bound)	p (Upper Bound)
A vs. C	7.9	<.0001	<.0001
A vs. P	13.8	<.0001	<.0001
C vs. P	6.9	<.0001	<.0001
L vs. M	1.4	.99	.99
L vs. R	3.5	.003	.003
M vs. R	2.0	.27	.27

Probability values in (B) were Bonferroni's corrected for six comparisons; *t* tests were two-tailed. A = Anterior; C = central; P = posterior; L = left; M = midline; R = right.