**Abstract**

**Review of Literature**

Study 1

Study 2

**Research Questions**

**Methodology**

Participants

Materials

Procedure

Scoring & Coding

**Results**

Study 1

Study 2

**Proposed Changes, Discussion & Conclusion**

**References**

**Review of Literature**

**Study 1**

**Cognitive control and linguistic ambiguity resolution**

**The bilingual advantage: Bilinguals & cognitive control**

**Study 2**

**The language learner: Proficiency and cognitive control**

**Non-natives and ambiguity processing**

**Research Questions**

Study 1

1. Does cognitive control engagement affect latency in processing temporarily ambiguous structures in early bilinguals?

Study 2

1. Does proficiency affect latency in processing temporarily ambiguous sentences in second language learners?
2. Does cognitive control engagement affect latency in processing the non-dominant language in second language learners?
   1. If so, does cognitive control engagement affect resolution of the ambiguity in temporarily ambiguous sentences?

**Method**

Participants

Pilot Study 1

Participants in Pilot Study 1 were 5 early bilinguals of Spanish and Catalan (3 female, 2 male). Four of the participants (Participants 2-5) were simultaneous Spanish-Catalan bilinguals, who began learning Spanish and Catalan before age 2. The fifth participant (Participant 1) was a sequential bilingual, who began speaking Catalan at the age of 7 when her family moved from Colombia, and therefore was factored into analyses as a separate profile. The participants were recruited from the Universitat Pompeu Fabra in Barcelona, Spain. Participants had completed between 1 and 4 years of university study. The average age was 20.6 (SD: 1.34).

The participants were asked to self-report their abilities in Spanish and Catalan on a scale of 0-5, where 0 was defined as “poor” and 5 as “near-native or native”. They ranked their abilities in the following five categories: reading, writing, speaking, listening and vocabulary knowledge. The four simultaneous bilinguals self-reported each of these abilities in Spanish and Catalan at near-native or native levels (M = 5, SD = 0); the sequential bilingual reported her Catalan at “near-native or native” for reading, and at “excellent” for the other four abilities. However, she also rated her Spanish listening and speaking abilities as “excellent”, and only ranked her reading, writing and vocabulary knowledge at the “near-native or native” level.

Participants were also experienced foreign language learners. The four simultaneous bilinguals reported the age at which they began learning English before age 13 (M = 7.5, SD = 3.27). The sequential bilingual reported learning English between 13 and 17. Their reported English abilities are listed in TABLE XXXXXX. Three participants (2, 3, 4) also began learning a fourth language after age 5 before age 13 (Japanese, German, French, respectively).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | 3 | 3.2 | 2.8 | 2.4 |
| 1 | 1 | 1.483239697 | 1.095445115 | 1.140175425 |

Pilot Study 2

Participants in Pilot Study 2 were 18 classroom learners of Spanish (7 female, 11 male). They were recruited from one of three Spanish language courses at Georgetown University (Intermediate I [3 participants], Advanced I [12], or Oral Review [3], a course that follows Advanced II for these students). Three early bilinguals (2 Korean-English and 1 English-Spanish bilinguals) were removed from the sample. The remaining 15 participants all identified themselves as native speakers of English only, all born in the United States, and with a mean age of 18.6 years (SD = 0.74). One participant began learning Spanish before age 10, attending a bilingual English-Spanish pre-school and kindergarten. Seven participants began learning Spanish between the ages of 10 and 13, and seven between 13 and 18. Spanish was the first classroom-learned language for 13 of these participants; one participant studied Hebrew and one studied French after age 13.

In addition to low accrual numbers in the Intermediate I and Oral Review groups, two intermediate students did not attain the minimum comprehension threshold (discussed below) and one Oral Review student was removed from the sample as an early bilingual, so analyses discussed beyond this point will only include the ten remaining advanced students.

**Materials**

Experimental Task

The design of the experimental task included two sets of stimuli with a [+/-conflict] dichotomy: the non-linguistic flanker task and the linguistic self-paced reading stimuli. The critical stimuli were interleaved such that each critical linguistic stimulus was preceded by a congruent or incongruent flanker trial.

*Non-linguistic Stimuli*

The flanker task asks participants to respond by pressing one of two buttons, according to the direction indicated by the center arrow on the screen in each trial. Each stimulus presents either one or five arrows in the center of the screen. Stimuli fall into one of three conditions: single center arrow, congruent center arrow, and incongruent center arrow. The congruent condition includes five arrows all pointing in the same direction, while the incongruent condition includes one center arrow pointing in one direction, flanked by two arrows on each side that point in the opposite direction of the center arrow. The critical comparison is between the congruent and incongruent stimuli. The direction of the center arrow is balanced across trials. Each stimulus is preceded by a fixation cross on screen for 1000 milliseconds (ms). Participants have up to 2000 ms to respond. Failure to respond within that time limit is scored as in an incorrect response.

*Linguistic Stimuli*

The linguistic stimuli were sentences presented via the self-paced reading moving window paradigm. Before each sentence was presented, a fixation cross lasting 1000 ms reset the participant’s gaze on the center of the screen to prepare the participant for the trial. Each stimulus consisted of a single sentence with each word masked by hyphens, presented on screen in its entirety. Participants uncovered one word at a time by clicking the space bar. As one word was uncovered, the previous word was re-covered by hyphens. The latency of reading times of each word was recorded by PsychoPy. Following each sentence, participants answered a true/false comprehension question (CQ).

The critical linguistic stimuli presented in the self-paced reading trials presented subject-object ambiguities, all of which are listed in the Appendix. These stimuli modulated transitivity of the verb in a preposed adverbial clause, as can be seen in (XXXXX) below (with each region of interest indicated by a backslash).

(XXXX) a. *Transitive*: Cuando el escultor \ acabó \ la obra \ tenía tres metros \ de altura.

b. *Intransitive*: Cuando el escultor \ volvió \ la obra \ tenía tres metros \ de altura.

Region: 1 \ 2 \ 3 \ 4 \ 5

“When the sculptor\ {finished/came back} \ the piece \ was ten feet \ in height.”

*Experimental Design*

Other than minor changes to accommodate a lower proficiency level (addressed below), the transitive (ambiguous) sentences were taken directly from Jegerski (2012). However, because the preceding flanker trial introduced an additional variable, the design departed from Jegerski’s (2012) original design. Two lists of flanker-sentence stimulus pairs were created in a Latin Square design. Each list was then divided into two blocks to counterbalance presentational order across participants. The 20 intransitive (unambiguous) sentences were distributed across these four blocks (five unambiguous sentences per block), always preceded by an incongruent flanker trial. Thus, each participant would complete one list, consisting of two blocks with a total of 10 flanker-sentence pairs of the [+conflict, +ambiguity] condition, 10 pairs of the [-conflict, +ambiguity] condition and 10 pairs of the [+conflict, -ambiguity] condition.[[1]](#footnote-1) Because the change to Jegerski’s (2012) original design meant participants would see both the transitive and the intransitive sentence, minimal lexical changes were made to the intransitive sentences in regions 4 and 5 so that participants were not reading the same sentences twice, but replacement words were chosen according to number of syllables and frequency.

In addition to the 30 critical flanker-sentence pairs addressed above, participants also completed 90 distractor flanker-sentence pairs and 30 unpaired distractor flankers to disrupt the participants expected patterns. The sentential distractor items included preposed adverbial phrases that included a transitive verb with an explicit complement, permanently ambiguous sentences and sequential prepositional phrases with high-low attachment ambiguity. The flanker distractor items were chosen such that each of the six conditions (right vs. left; congruent vs. incongruent vs single) was balanced across all trials.

Frequency of the words in the linguistic stimuli and distractors was assessed using the *Corpus del Español*’s WordAndPhrase tool, which compiles 20 million words from texts from the 1900s and 2 billion words from texts published online since 2014. Words in the stimuli that were not among the 3000 most frequent words on the corpus’ list were included in a decontextualized vocabulary test given to Beginner 1 students to assess how well more advanced students would recognize these words. The students in Beginner 1 were asked to translate each word and rate how well they recognized the word on a scale of 1-4. If words in this test met three of the following four criteria, they were included in the critical stimuli: (i) the word is covered in the vocabulary lists of Beginner 1 or 2 (i.e. Intermediate I students are expected to know this word); (ii) the word elicited an exact translation accuracy of 85% or higher; (iii) the word received a correct recognition rate among all raters of at least 3; and (iv) the word is a cognate with no more than two graphemic changes due to Spanish phonotactics and spelling norms (e.g. violín, escultor). The seventeen words that did not meet these criteria were included in a vocabulary training, which contained a total of 42 words. This training is addressed below.

COMPREHENSION QUESTIONS

Vocabulary Training

Study 2 participants were all required to complete a vocabulary training on PsychoPy before completing the study to establish a baseline vocabulary. This training presented 42 vocabulary items used in the stimuli, 17 of which appeared in the critical stimuli. The 42 words were divided into two blocks. The words in each set were presented one at a time with a photo that represented the word and a translation in English. After 1000 ms, participants were invited to move to the next word if and when they were ready. After each set of 21 words was presented, participants had an opportunity to review the words again or take a quiz to continue. The quiz presented four of the photos from the set and asked participants to identify which photo represented the word shown on the screen. A minimum score of 85% was required on each quiz in order to continue to the next phase of the study.

Language Background Questionnaire

All participants completed a language background questionnaire (LBQ), edited from the LEAP-Q (Marian, Blumenfeld & Kaushanskaya, 2007). Information regarding languages spoken was collected for all participants, but the participants in Study 1 completed detailed questions for Spanish, Catalan and English; participants in Study 2 completed detailed questions for English and Spanish. The questions were presented via computer.

Elicited Imitation Task

To complement institutional placement in Study 2, the Elicited Imitation Task (EIT) was used as a secondary proficiency measure (SEE APPENDIX XXXXX). The EIT has been validated as a reliable measure of global proficiency as compared to the Simulated Oral Proficiency Interview (SOPI), a global proficiency measure based on the American Council of the Teaching of Foreign Languages proficiency guidelines (Ortega, 2000; Ortega et al., 2002). In this task, participants are recorded listening to and repeating a set of 30 pre-recorded sentences. These sentences become progressively more complex structurally. The recordings are then scored on a scale of 0-4 according to the accuracy of each repetition, and the total score for each participant falls between 0 and 120 (see APPENDIX XXXXXX).

Items changed from Jegerski:

* Después de que empezó el maratón/después de que corrió el maratón 🡪 “empezar” now “llegar”
* acabar la obra 🡪 terminar la obra
* 3 metros de altura 🡪 diez pies de altura
* Cinco días después 🡪 antes de que
* Concursante 🡪 participante
* Cuando la novia descendió la escalera le pareció muy larga 🡪 Cuando la chica lava los platos se quedan bien sucios
* Limpiecita 🡪 muy limpia
* De una vez 🡪 inmediatamente

1Counted, “separately”

2Counted

3Counted

4Counted

5Counted, flankers very slow

6 counted

7 out – Daniel Arenas, native speaker

8 out – Stephen Cho, Korean, flankers residuals = negative

9 counted

10 counted

11 – oral review

12 - counted

13 - counted

14 – advanced, not counted after second calculo

15 - counted

16 – advanced, .633

17 - counted

18 – oral review (.633)

19 – intermediate (70%, but weird data)

20 – intermediate (only remaining intermediate)

21 – intermediate (.60),

22 – data lost, oral review, Korean speaker

23 - counted

https://journals.sagepub.com/doi/abs/10.1177/1367006917709097

|  |
| --- |
| 12.88888889 |
| 2.643279801 |

**Procedure**

Study 1

Early bilinguals in Study 1 were recruited from a group of students at the *Universitat Pompeu Fabra* who participated in a Spanish conversation program with Georgetown’s summer abroad program in Barcelona. They were paid 10€ for completing the study. The entirety of the study was conducted in Spanish. Before beginning the study, the researcher explained the study broadly in a quiet room, obtained informed consent and assigned participants to one of four groups. These groups determined the list and presentational order of blocks for each participant. Following this, participants completed three training phases to prepare for the experimental phase: isolated flankers, isolated sentence readings, and interleaved flanker-sentence reading pairs. Participants needed to achieve minimally 80% accuracy in order to move on to the experimental phase. Once they completed the training phases, they had an opportunity to rest and ask any questions before beginning their two experimental blocks, each of which contained 60 flanker-sentence pairs, 15 unpaired flankers, and a brief break. Following the first block, participants had another opportunity to rest before completing the second block. In addition, if the participant answered three consecutive flankers or three consecutive CQs incorrectly, they were forced to take a 10-second break, during which they were reminded that both speed and accuracy were important to finish the study in a timely fashion. Upon completion of the experimental component of the study, participants completed the LBQ.

Study 2

Emergent bilingual language learners in Study 2 were recruited from Spanish language classes at Georgetown University. They were rewarded 2 points on a mid-semester test for participating in the study. Participants completed the study in a large computer lab with up to three other participants. These participants all wore headphones and were sat in opposite corners of the lab facing away from each other to minimize interference. Before beginning the study, the researcher explained the study broadly, obtained informed consent and assigned participants to one of the four groups to determine list and block order. Participants completed three training phases to prepare for the experimental phase: isolated flankers, isolated sentence readings, and interleaved flanker-sentence reading pairs. The explanation of the study, informed consent, and first two training phases were conducted in English. The instructions to the interleaved training phase were presented in English, but then participants were instructed that the actual stimuli and CQs would be presented in Spanish. This was done to facilitate the transition to Spanish and initiate the so-called “monolingual mode” in Spanish. All computerized instructions and stimuli following this point were presented in Spanish. In order to move onto the experimental phase, participants needed to achieve 80% accuracy or higher on each practice phase.

Once they completed the training phases, they had an opportunity to rest and ask any questions before beginning their two experimental blocks (identical to the blocks in Study 1). Like in Study 1, participants who answered incorrectly on three consecutive flankers or CQs were forced to take a short break. Upon completion, participants completed the LBQ, after which they were escorted from the laboratory to complete the EIT, the final component of Study 2, in another space to control the noise level in the laboratory.

Scoring & Coding

*Self-paced Reading*

PsychoPy latency data collection occurs at the word level, so regional-level latency was calculated for the 40 critical stimuli (30 per participant).

*Outliers*

In order to account for data that was likely affected by external factors, data was trimmed at the group and individual level. At the group level, a maximum of 6000 ms (Jackson, 2010 – this is from JEGERSKI’s book) and a minimum of 100 ms (Luce, 1986 ---Jegerski’s book) was set. Values outside of these bounds were replaced with the limit. A total of 6 values exceeded 6000 ms and 0 values fell below 100 ms, accounting for 0.2% and 0% of the data. At the individual level, maximums and minimums were established for each participant at +/- 2.5 standard deviations (SDs) from the mean RT, and these individual limits replaced any value that fell outside of the bounds (Marijuan, dissertation). A total of 66 values, including the 6 mentioned above, were replaced with the participants’ upper bound, accounting for 2.3% of the data. This falls within the standard recommended by Jegerski (2014).

*Comprehension Check*

After removing outliers, the data was then cleaned with regards to comprehension checks, with a threshold set at 60% accuracy on the CQs (Marijuan, dissertation). All participants met this threshold (5 early bilinguals in Study 1; 14 emergent bilinguals in Study 2). At the item level, sentence-flanker pairs with incorrect CQs were removed from analyses. This accounted for the removal of 6.67% of early bilingual data (10 stimuli pairs distributed across the 5 participants) and 26.19% of emergent bilingual data (110 stimuli pairs across the 14 participants). This left a total of 140 flanker-sentence stimuli pairs for the early bilinguals and 310 pairs for the emergent bilinguals.

*Residual Reaction Times*

At this point, residual RTs (RRTs) were calculated to normalize the data, using the RTs of the sentences that the participants comprehended and that complied with the above established criteria. This process requires plotting coordinate pairs of raw RTs and the length of the corresponding regions. The line of best fit is then calculated for each participant. The slope of this line reflects the average reading time at different word lengths, and the y-intercept represents the baseline reading time for each word, once differences of length are factored out. This information is used to generate an expected RT for a region based on the number of letters in that region. The expected RT is subtracted from the actual reading time, resulting in the RRT. Positive RRTs reflect a region read slower than expected; negative RRTs reflect regions read faster than expected. The aggregate RRT means for each analogous region were then calculated.

*EIT Scoring*

The elicited imitation test includes 30 sentences of increasing length and complexity, each of which is scored from 0 to 4 (Ortega, 2000). A research assistant scored the recordings of the EITs. Before scoring the participants from this study, the researcher and research assistant independently scored two EITs collected for a previous study. After this first round, Cohen’s kappa (κ) was run to determine the extent of agreement between the two coders. Following recommendations from Cohen (1960) and Lowry (2019), linear weighting was used to account for the fact that the scores awarded were ordinal (increasing scores represent increasingly more successful production of the target), and the reported kappa is observed as a proportion of the maximum possible kappa given the marginal frequencies (this is recommended in cases when raters are not limited to a certain number of ratings per category). Concordance was moderate between the two raters, κ w = 0.7024 (95% CI, 0.196 to 0.521), SE = 0.0622. Disagreements were discussed one by one until a consensus was reached. After this, the two raters individually scored recordings from the first two participants from Study 2. Cohen’s Kappa was calculated again and a much higher rate of agreement was found, κ w = 0.9472 (95% CI, 0.8091 to 0.9463), SE = 0.035. This is considered excellent by Fleiss (1981) and almost perfect agreement by Landis and Koch (1977). Following the second round of independent scoring, the research assistant scored the remainder of Study 2 participants’ recordings. ~~Because one of the three participants from the highest level of classroom Spanish was a native bilingual, only two remained in this group. Following the analyses of the EIT scores, one of these two Oral Review students had a EIT score lower than most students in the advanced level and one of the three students from the intermediate level. Because removing this participant would result in t was determined to exclude Oral Review from analyses.~~

EIT FROM ARIEL

The test produces a score on a scale from 0 to 120 (see Appendix D for scoring protocol). Due to the subjective nature of the scoring protocol, the EIT was coded by three independent raters. To ensure interrater reliability, first, each rater scored the responses from the same three EITs using the scoring protocol from Ortega (2000) (see Appendix D). The raters then compared their scores and, where discrepancies arose, arguments for choosing the scores were discussed, scoring was clarified, and a consensus was made as to the most appropriate score. Based on the results from this process, three more EITs were chosen and each rater scored them. Once again, results were compared and interrater reliability was calculated by totaling the number of items where discrepancy arose divided by the total number of items, reaching an interrater reliability score of 86.67%. Table 3 presents the descriptive statistics for EIT scores for the four initial recruitment groups, second-semester, fourth-semester, sixth-semester, and advanced. Three participants, all from second-semester, were removed from the study due to lack of an EIT score after their answers failed to be recorded, one participant was removed from the advanced group for being a heritage learner, and one participant was removed from the sixth-semester group as she reported a language other than English as her L1 and did not begin to study English until high school. This yielded a total L2 learner group of 90 participants. A one-way ANOVA and subsequent post hoc Scheffé showed a significant difference between all four groups on the EIT, F(3,83) = 84.71, p = .000 (η2 = .76, observed power = 1.00), with a large effect size between groups15 . Due to the overlap between the large ranges of scores for L2 groups, participants from the original recruitment were redistributed into four new groups based on percentile scores for the EIT. Four evenly distributed groups of EIT scores were established based on the 25th (n=22, range 10-45), 50th (n=23, range 46-83), 75th (n=23, range 84-111), and 100th (n=22, range 112- 120) percentiles. A one-way ANOVA showed that there was a significant difference by percentile group, F(3,83)=393.24, p=.000, partial ŋ2 =0.93, with a large effect size, and the post hoc Scheffé test indicated that all four percentile groups performed significantly different. In order to create a clear distinction in proficiency between each group, group boundaries were established so that the highest score of one group was not within one point of the lowest score of the next group, using the percentile group boundaries as a starting point. This yielded the range of proficiency scores for the four L2 learner groups presented in Table 4. A one-way ANOVA found a significant effect by group for the EIT scores, F(4,119) = 637.57, p=.000, partial ŋ2 =0.95, with a large effect size. A post hoc Scheffé test indicated that all four participant groups are significantly different from one another, but the high advanced group was not significantly different from total possible score, as indicated with indices in Table 4, where non-matching indices indicate significant differences at p<0.05.

Results & Analyses

*Study 1*

*Comprehension*

For the CQs that followed each stimulus, early bilinguals responded correctly on 93.33% of the questions (Range: 90.00% to 96.67%). In terms of transitivity, early bilinguals responded correctly to CQs for 92.00% of transitive sentences and for 96.00% of intransitive sentences.

*Latency in reading times*

Separate repeated-measures analyses of variance (ANOVAs) were conducted for the regions of interest of transitivity, following Jegerski (2012): Regions 3, 4 and 5. A repeated-measures ANOVA was also conducted for Region 1 to assess whether the preceding non-linguistic conflict played a role in the first region of the linguistic trial. Because the pilot study only included unambiguous sentences that were preceded by incongruent flanker trials, the analyses included one independent variable, trial condition, with three levels ([+ambiguous, +incongruent], [+ambiguous, +congruent], [+unambiguous, +incongruent]). For the analyses by subjects (F1), subject was the random factor and the within-subjects factor was condition of the flanker-sentence pairs, as listed above. For the analyses by item (F2), item was the random factor and flanker-sentence condition was the within-subjects factor.

The repeated-measures ANOVAs, with a Greenhouse-Geisser correction when sphericity was not assumed, determined that participant mean RRTs did not differ significantly across the three conditions in any of the four regions analyzed. The repeated-measures ANOVAs also revealed that item means did not differ significantly across the three conditions in these four regions. These results are reported in Table XXXXX.

Table XXXXX

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | By subjects | | | | By items | | | |
|  | *df* | *F1* | *p* | *ηp²* | *df* | *F2* | *p* | *ηp²* |
| Region 1 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 1.038\*, 2 | 0.101 | 0.774 | 0.025 | 2, 17 | 0.037 | 0.963 | 0.002 |
| Region 3 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 1.010\*, 2 | 0.497 | 0.521 | 0.111 | 2, 17 | 0.228 | 0.638 | 0.012 |
| Region 4 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 2, 2 | 0.519 | 0.614 | 0.115 | 2, 17 | 0.705 | 0.471 | 0.036 |
| Region 5 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 2, 2 | 0.045 | 0.956 | 0.011 | 2, 17 | 0.411 | 0.666 | 0.021 |

\* Value corrected for sphericity with Greenhouse-Geisser correction

*Study 2*

*Comprehension*

For the CQs that followed each stimulus, emergent bilinguals from the three institutional levels responded correctly to 73.81% of the questions (Range: 60.00% to 93.33%). In terms of transitivity, emergent bilinguals responded correctly to CQs for 71.79% of transitive sentences and for 77.86% of intransitive sentences.

*EIT*

Because of the attrition of two

*Latency in reading times*

For the analyses by subjects (F1), separate repeated-measures ANCOVAs were conducted for the regions of interest of transitivity (regions 3, 4, 5). Region 1 was also considered in a separate analysis to assess whether the preceding trials’ congruency affected residual reading times at the beginning of the sentence stimuli. In these analyses, subject was the random factor and the within-subjects factor was condition of the flanker-sentence pairs ([+ambiguous, +incongruent], [+ambiguous, +congruent], [+unambiguous, +incongruent]). Given the overlap in EIT scores across institutional levels, EIT was used as the covariate to account for proficiency, instead of the course level. For the analyses by item (F2), repeated-measures ANOVAs were run for each of the four regions listed above, with item as the random factor and flanker-sentence condition as the within-subjects factor.

Table XXXXX

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ANCOVA by subjects | | | | ANOVA by items | | | |
|  | *df* | *F1* | *p* | *ηp²* | *df* | *F2* | *p* | *ηp²* |
| Region 1 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 2, 11 | 1.711 | .202 | .125 | 2, 15 | 1.516 | 0.234 | 0.082 |
| EIT Score x Ambiguity/Conflict | 2, 11 | 1.365 | .275 | .102 |  |  |  |  |
| Region 3 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 2, 11 | 1.637 | .216 | .120 | 1.370\*, 15 | 0.743 | 0.483 | 0.042 |
| EIT Score x Ambiguity/Conflict | 2, 11 | 1.572 | .228 | .116 |  |  |  |  |
| Region 4 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 1.243\*, 11 | 1.011 | .350 | .078 | 2, 15 | 1.559 | 0.225 | 0.084 |
| EIT Score x Ambiguity/Conflict | 1.243\*, 11 | .289 | .648 | .024 |  |  |  |  |
| Region 5 |  |  |  |  |  |  |  |  |
| Ambiguity/Conflict | 2, 11 | .704 | .505 | .055 | 1.094\*, 15 | 0.250 | 0.645 | 0.014 |
| EIT Score x Ambiguity/Conflict | 2, 11 | .698 | .507 | .055 |  |  |  |  |

\* Value corrected for sphericity with Greenhouse-Geisser correction

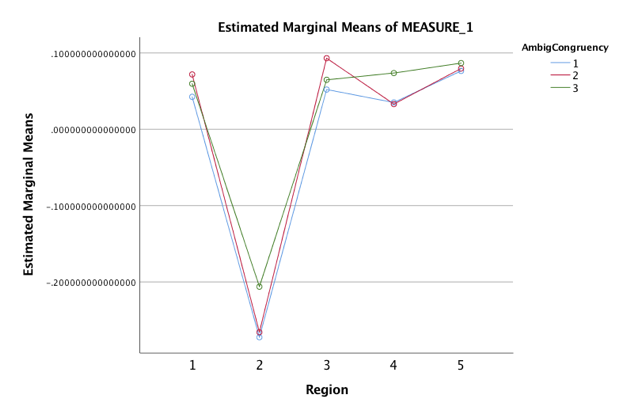
Proposed Changes, Discussion & Conclusion

*Study 1*

For the pilot study with early bilinguals, only five participants were recruited to test the materials, so the statistical power was limited. However, the data collected from these five participants is still informative. In terms of transitivity, we see a slow-down at Region 2 in the intransitive (unambiguous) sentence. This was not found in Jegerski’s (2012) results, so this is an unexpected result. While there were differences in stimuli presentation and data handling (word-by-word vs region-by-region; raw RTs vs RRTs), these are unlikely to have created this difference since Region 2 was only one word long and the RRTs are calculated using the number of letters, which is balanced across conditions. It may reflect a the transitivity of the words: transitive verbs in conditions 1 and 2 accept direct objects, but the subcategorization matrix of an intransitive verb slows the reader down, essentially to insert a pause where one might expect a comma in a non-experimental setting. Upon more fine-grained analysis of individual participants, it doesn’t localize in one participant, so more data is needed to see if this is a generalizable finding.

In terms of congruency of the preceding non-linguistic trials, the results reveal the anticipated pattern: activation of cognitive control would facilitate the resolution of the ambiguity in the critical region of the sentence (Region 3). The means of the two ambiguous conditions (regardless of the congruency of the preceding stimulus) are effectively indistinguishable in regions 2, 4 and 5. It is important to remember when considering these results that the sentences in these two conditions were counterbalanced, so while some participants saw sentences 1-10 with an incongruent preceding flanker, others saw these same sentences with a congruent preceding flanker, so the differences at Region 1 and Region 3 can be reasonably associated with the preceding trials’ congruency: incongruence triggers faster initial reading and faster resolution of the ambiguity.

Comparing the three conditions simultaneously is more difficult because the study is missing the fourth permutation of the two variables (transitivity and congruency). Does activation of cognitive control allow the parser to resolve ambiguities faster than to process unambiguous structures (Region 3, Condition 1 vs Condition 3)? This seems unlikely, but ushers in the importance of the inclusion of the [-ambiguous, -congruent] condition in the full study that will follow, in order to compare both variables independently as they should be.

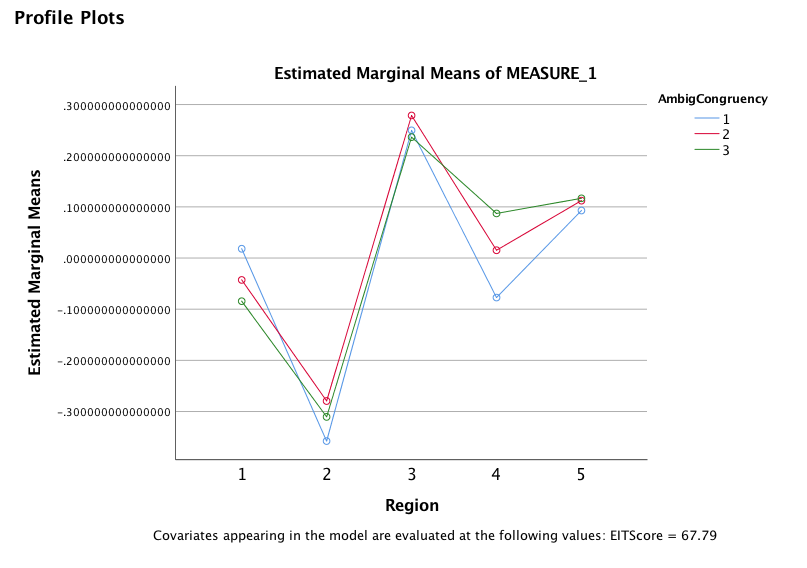


*Study 2*

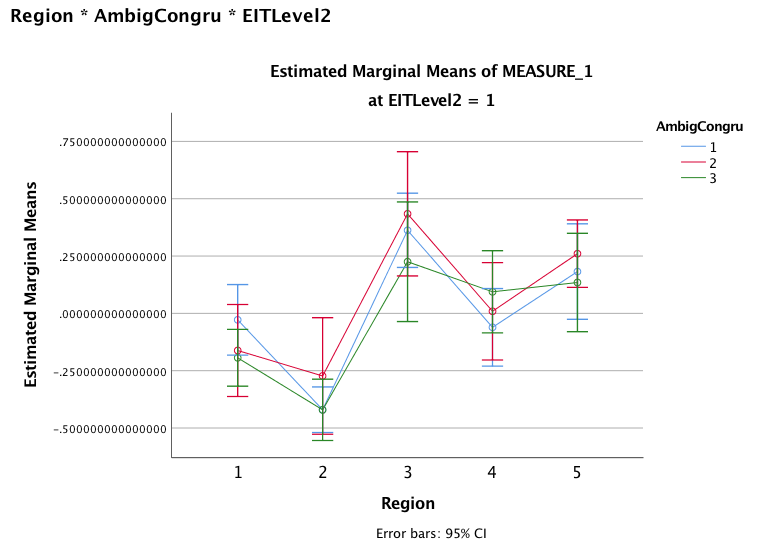
It is difficult to answer the research questions as posed for Study 2. Firstly, in two of the three institutional levels, only three participants were recruited. Secondly, with so few recruited, auto-selection bias plays a bigger role, and the high proficiency participants in one group were higher than the low-proficiency participants in the next group up. However, the secondary proficiency measure used provides a continuous variable that could be used as a covariate with latency measures.

Like in Study 1, while Condition 3’s stimuli were balanced for word length and frequency, Condition 1 and 2 are the exact same stimuli with differing conflict [+ or -] in the preceding trial. Therefore, we can draw stronger conclusions from these initial results in these two conditions. Interpreting the plot of residual reading times for Study 2, we see again that Condition 2 is the slower condition at Region 3, but the three conditions seem negligibly different. Meanwhile, the conditions do diverge more at Region 4. This may indicate a delayed processing of the ambiguity in second language learners that we do not see in native speakers (i.e. early bilinguals in Study 1). If this spillover effect is the manifestation of this L2 delayed processing, the differences between Condition 1 and 2 in Region 4 coincide with the results of Study 1. Of course, without more data from the three levels, the hypotheses cannot be confirmed or rejected.

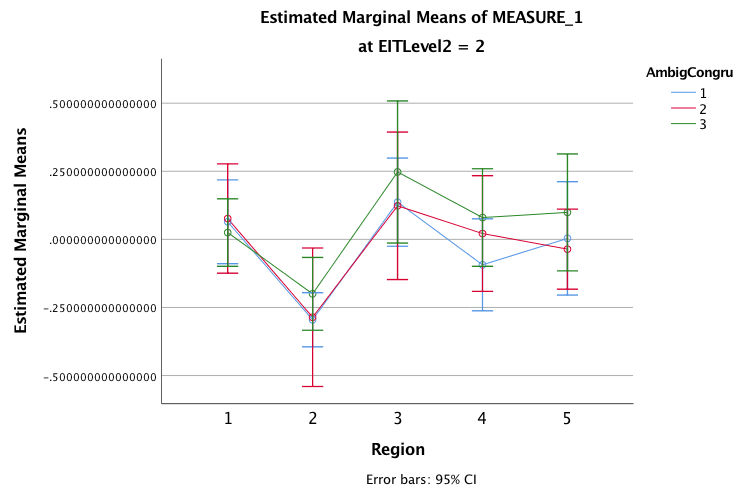
There may be a refractory period for the learners following the resolution of conflict which we can see in Condition 1 that we don’t see in Study 1. While Jegerski (2012) only considered the critical region and spillover regions, this study (in its full rendition) will likely have to consider Regions 1 and 2 because of the proximity to the non-linguistic task and the spillover effects that that flanker and Stroop tasks have.



A post-hoc division of the group according to EIT score (using the median EIT score as the dividing point) revealed a trait that would align with two of the expected patterns. In the lower proficiency group, means were slightly slower in Condition 2, but there the difference between Condition 1 and 2 was generally consistent across all regions except Region 1.



Meanwhile, in the higher proficiency group, we can observe highly similar means in the two ambiguous conditions, regardless of cognitive control engagement, except at Regions 4 and 5. While we cannot make any conclusions without more data to narrow the error rate, the current results might reflect the two roles of cognitive control in bilingual language processing: at lower proficiencies, cognitive control is working to suppress the dominant language, but at higher proficiencies, cognitive control is not recruited for this purpose but rather to resolve higher-order problems like ambiguity resolution. While this cannot be concluded from the present data, it does present rationale to continue recruiting lower-proficiency learners (i.e. intermediate students) despite their low comprehension rates.



Pearson’s correlation between proficiency (EIT Score) and comprehension check:

Two-tailed, with P23: *r =* 0.331, *p =* 0.247

Two-tailed, without P23: *r* = 0.634 , *p* = 0.023

Expected patterns were depletion, maximal engagement and heightened engagement, but the pattern observed in this post-hoc analysis does not support any of these patterns. One

1. Depletion: Given the recruitment of the LIFG and other prefrontal structures observed by Abutalebi (2008) during the processing of a non-proficient L2, the demands of using an L2 at these lower proficiency levels may result in a carryover effect, a refractory period during which cognitive control resources available to the participants are effectively temporarily depleted by the high cognitive load. This would relate to the results obtained in child processing by Huang and colleagues (2016). An interaction would be observed between Stroop congruency on trial *n*-1 and proficiency level, where higher proficiencies who do not rely so heavily on cognitive control for processing receive a benefit from cognitive control activation, while lower proficiencies process the ambiguity slower when the *n*-1 trial is conflictive.
2. Maximal Engagement:
3. *“Maximally Engaged” Cognitive Control.* This alternate possible set of results assumes that cognitive control of low proficiency participants is engaged throughout the study due to the use of the L2, and so preceding incongruent Stroop trials would not result in improved recovery from misanalysis for these participants. Therefore, hypothetically, an interaction would be observed between Stroop congruency on trial *n*-1 and proficiency level, as the high proficiency groups would perform better following incongruent Stroop trials, as depicted in Figure 5, while the low proficiency groups would not perform differently under the two conditions.
4. Heightened Engagement:
5. *“Heightened Engagement” of Cognitive Control.* This final alternative would show that cognitive control engagement can be manipulated to aid recovery from misinterpretation for each proficiency level (See Fig. 6). This finding would suggest that the increased consideration of the incorrect goal in Pozzan and Trueswell (2015) is a reflection of proficiency level and not overloaded cognitive control. Each proficiency level would respond to the cross-task manipulation similarly, and so perform similarly under both ambiguous conditions.

Changes to be made:

1. 4 conditions (no conflict, unambiguous)
2. resolve issue with unambiguous sentences
3. learners: intermediate 2?; oral review?; post-study abroad (or graduate students)

**References**

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, 20, 37-46.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley. [ISBN](https://en.m.wikipedia.org/wiki/International_Standard_Book_Number) [0-471-26370-2](https://en.m.wikipedia.org/wiki/Special:BookSources/0-471-26370-2).

Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. 33 (1): 159–174.

Cited

(Frazier & Fodor, 1978;

Frazier & Clifton, 1998;

MacDonald Pearlmutter, Seidenburg, 1994;

Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995).

Lowry (Vassarstats – <http://vassarstats.net/>)

(Thompson-Schill, Bedny, & Goldberg, 2005;

Novick, Kan, Trueswell & Thompson-Schill, 2009;

January, Trueswell, Thompson-Schill, 2009,

Teubner-Rhodes et al., 2016).

UNSURE????

All other citations added to refme

Gratton, Coles, & Donchin, 1992

January, Trueswell & Thompson-Schill, 2009

Kroll & Bialystok, 2013;

Rosselli, Ardila, Araujo, Weeks, Caracciolo, Padilla, Ostroski-Solis, 2000

Ullsperger, Bylsma & Botvinick, 2005

Ye & Zhou, 2009

1. Proposed Change: The decision to present all unambiguous stimuli after a [+conflict] flanker was made to allow for stronger comparisons and to avoid attention to the similar sentence structures between the transitive and intransitive sentences, but following data collection, it became clear that a stronger study would also include the fourth permutation of the conditions: [-conflict, -ambiguity]. The proposed change would mean both lists would include each unambiguous sentence, and the preceding flanker condition would be counterbalanced across the two lists. [↑](#footnote-ref-1)