# Final Critique of Cotter & Ferreira (2024), "The relationship between working memory capacity, bilingualism, and ambiguous relative clause attachment"

Jennifer A. Sheridan

Notre Dame de Namur University

CPY 4896: Research Methods and Proposal

Dr. Michael Drexler

December 2, 2024

# **Final Critique of Cotter & Ferreira (2024), "The relationship between working memory capacity, bilingualism, and ambiguous relative clause attachment"**

This study investigated the impact of bilingualism and working memory on an individual's ability to resolve syntactic ambiguity and assesses how different language backgrounds might have different results. Language can sometimes be globally ambiguous—the authors used the example sentence, "The man saw someone on the hill with a telescope." Did the man use a telescope to see someone on the hill (attaching the relative clause to the first noun or N1)? Or did the person on the hill have a telescope (attaching the relative clause to the second noun or N2)? Research has shown that people whose first languages (L1) are English, Arabic, Basque, Mandarin, or Swedish tend to attach the relative clause to N2, which is frequently called a low attachment preference, whereas native Spanish, Dutch, French, German, and Russian speakers are more likely to have a high attachment preference (attaching the relative clause to N1). These preferences are mild, ranging from 55 to 65%, but statistically significant.

Swets et al. (2007) found that a higher working memory span leads to a preference for low attachment. This result was different from earlier research, so the authors wanted to replicate this study while also expanding the language groups used to more clearly understand the relationships between the variables. They predicted that bilinguals more proficient in their second language (English) would attach relative clauses in their first language the same way they do in their second, stronger language. There were three notable results: the relative clause attachment task revealed clear cross-linguistic preferences, consistent with much of the literature on differences in parsing preferences across languages. The L1 Mandarin-L2 English and L1 English groups preferred low attachment; the L1 Spanish- L2 English groups preferred high attachment. For relationship between working memory and relative clause attachment preferences, the participants with larger working memory spans preferred low attachment more than those with lower working memory spans, replicating the Swets et al. (2007) results. Finally, language proficiency did not show a significant association with the attachment preference of both bilingual samples – the relative clause attachment preferences were not related to language experience – which was a novel result.

## Design and Variables

This is a correlational study design. The independent variable (IV) is the different language groups (Spanish-English bilinguals, Mandarin-English bilinguals, and English monolinguals). The study examines two dependent variables (DVs): working memory span scores (DV1) and relative clause attachment preference (high vs. low; DV2). By analyzing both DVs, the researchers were able to explore relationships between working memory capacity, attachment preferences, and language group, as well as investigate potential bidirectional or mediating effects.

## Participants

Participants were from the undergraduate student recruitment pool at UC Davis and received course credits for participation. They included 338 adults: 218 identified as female, 120 as male, with a median age of 20.1 years. Spanish tends to have a high attachment preference while both English and Mandarin have a low attachment preference, so participants were broken into three groups: L1 Spanish-L2 English bilinguals (contrasting attachment; 118 participants), L1 Mandarin-L2 English bilinguals (complementary attachment; 105 participants), and L1 English monolinguals (115 participants).

All bilingual participants considered English as their second language, and the monolingual participants were not significantly exposed to another language. The Spanish and Mandarin bilingual groups differed in years of residency in the US, current language exposure, and first and second language proficiencies—L1 and L2 immersion percentages are averages of participant's score calculated from their age, age of acquisition, and years of language use. Spanish speakers were slightly more proficient in English than Spanish; Mandarin speakers were much more proficient in Mandarin than English, with differences especially striking in writing skills. They determined that no participants had learning disabilities such as dyslexia that may have affected their results.

Method

As the study ran October 2020 through April 2021, it needed to be online of necessity; data was collected using Qualtrics. Participants were given four tasks via the online software; all tasks and instructions were given in L1 (the participant's native language). This study was a modified replication of Swets et al. (2007, Study 1); all three tasks from that study were included here in a modified form. Participants completed two working memory span tasks (reading and spatial span); a relative clause attachment task; and 6-question, modified version of the Language History Questionnaire (LHQ) 3.0.

The working memory span tasks were identical across all three language groups regardless of their first language. Items were scored as correct if both parts of a task were correct (both processing and recall for each item); the number of correct items of out 36 items for each task type became the score. For the reading span task, the participants were given sets of 3, 4, 5, or 6 items. Each item consisted of a sentence (e.g. The pilot flew the plane to the store on Halloween) followed by a question mark (there to remind them to answer if the question made sense – yes or no) over a letter (e.g. J). After each set, participants were asked to recall the letters (like the "J") in the order they had been presented in the set. This was duration based, not reaction time based (i.e. 12 seconds for the 3-item set up through 24 seconds for the 6-item set).

For the spatial span task, each item consisted of a letter (G, F, or R) floating in the middle of a blob shape; half the presentations were normal and half were mirror-reversed, and they were rotated through 8 possible rotations. The item was shown for 5 seconds, then participants were asked whether the letter was shown in normal or mirrored orientation. The recall task at the end of each set was to draw the direction in which the top of each letter was initially shown (i.e. pointing straight up ↑ if it was at 0 degrees, canonically upright).

The relative clause attachment task was set up such that participants viewed the same list of 100 items in the same order—20 items were experimental, while 80 were neutral or unrelated fillers (including about 40% that were relative clauses that modified referents in the subject or object position). The experimental sentence structure followed the same order in each experimental item. For example, "The friend of the movie star who was sitting on the balcony was under investigation" = N1 followed by N2 followed by relative clause followed by matrix verb phrase. After each sentence, participants were asked a question designed to measure their attachment decision (i.e. Who was sitting on the balcony?) The two possible responses were N1 (the friend; high attachment) or N2 (the move star; low attachment). The responses were placed one above the other and counterbalanced so that N1 appeared above half the time and N2 appeared below half the time. The presentation time was consistent (the sentence appeared for 5 seconds, the question for 3 seconds, then a fixation cross for 500 ms, and finally the response options for 4 seconds) and each item was presented one at a time. The score on this task was the number of N1 or high attachments divided by 20 possible answers.

The language history questionnaire (subset of LHQ 3.0) was given to assess language proficiency. Participants answered questions about current age, country of residence, years residing in the US, age of language acquisition, self-evaluation of proficiency, and dialect spoken (like Mexican Spanish). Language proficiency was broken down by reading, writing, speaking and listening proficiencies on a scale of 1-7 where 1 is very poor and 7 is excellent.

## Construct validity

In Morling's rubric "Interrogating the Three Types of Claims Using the Four Big Validities," she notes that association claims are usually supported by correlational studies, as we have here. She begins with construct validity and asks that we look at how well the researcher has measured each of the variables in the association as we want to be sure that the study measures what it thinks it is measuring.

All tasks in this study were modeled after previous research, which helps maintain consistency with established methods, strengthening the likelihood that the tasks measure the intended constructs. When creating the working memory span tasks, the authors worked hard to ensure construct validity. For example, 42 participants' results were excluded due to low accuracy (below 80%) on the filler sentences. Filler sentences are frequently used in psycholinguistic experiments to make sure that any effect measured is due to the design of the experimental sentences and not due to chance or demand characteristics (i.e. the participant has figured out what is being looked for and responds accordingly). They can also keep participants more engaged and less fatigued by the task at hand.

For the relative clause attachment task, the 100 items presented to the participants were translated to be consistent across the three languages, however they could not control for sentence or word length since Chinese is character-based, not alphabet-based like English and Spanish. They did control sentence length between Spanish and English to ensure there was no greater demand on working memory. Each translation was double checked by native speakers to ensure words were commonly used. Additionally, they replaced gendered experimental sentences as this was not the purpose of this study and they did not want clauses to resolve or create ambiguity based on gender stereotypes. This all helps ensure that this task is really measuring attachment preference and not something else.

While this study does a thorough job with construct validity overall, perhaps the weakest link is with the language group variable. They did a good job with the language history questionnaire so they could use different aspects of language acquisition and proficiency as covariates in the analysis. However, using a self-report may not have been sufficient to ascertain language proficiency in both languages.

### Statistical validity

The researchers were very thorough in their approach to statistical validity. When determining the necessary size of the participant pool, they conducted a power analysis using the R package pwr, setting the power ($1 - \beta$) at 0.95 and $\alpha$ at 0.05. Based on previous studies employing similar tasks, which reported an effect size (odds ratio) of 1.47, they determined that a sample size of at least 75 participants would be needed to detect the effect. However, since their goal was to replicate a study, they aligned their participant numbers with Swets et al. (2007), still falling well within the acceptable range of participants per sample as determined by the power analysis.

In analyzing the working memory span results, they used an ANOVA to compare performance differences across the three samples, followed by a Bonferroni test (which adjusts the level of significance when making multiple comparisons, avoiding Type 1 errors), revealing a statistical significance. The preference for high attachment was significantly greater in the Spanish sample (*M* = .64, *SD* = .19) than in the Mandarin (*M* = .31, *SD* = .18) and English (*M* = .40, SD = .27) samples. That said, despite the fact that the Mandarin and English samples both preferred low attachment, the English low attachment ended up being significantly greater than that of the Mandarin sample ($F(2, 335) = 59.59$, $p < .01$). The Bonferroni test clarifies that this statistical significance is not due to chance.

The three samples differed significantly in their self-reported L1 proficiency, which would have been a stronger result if there had been an objective measure utilized as well. The English sample was monolingual, so they used a one-way ANONVA between the Mandarin and Spanish samples, finding that the Spanish sample had significantly higher L2 proficiency than the Mandarin sample. Notably, the reading and spatial span scores did not significantly differ across samples.

The correlation analysis found that the participants were generally very consistent in their attachment preferences (if they preferred high attachment for one item they generally preferred high attachment for the other items), which was similar to Swets et al. (2007). Looking at Cronbach's alpha across the measures, they all had moderate to high internal consistency reliability (>.70), and then the reliability value for attachment preference was relatively high as well. In the correlations for all three language groups, both working memory measures were significantly and negatively correlated with high attachment, which was again similar to Swets et al. (2007) although slightly weaker.

The researchers used binomial logistic regression to examine the relationship between working memory (WM) span and attachment preference across the three language samples, as the dependent variable (attachment preference) was dichotomous (high or low). This method was chosen because it allows for the inclusion of continuous predictors, such as WM span, L1 proficiency, and L2 proficiency, to assess how these variables influence the likelihood of high versus low attachment preference. They combined the reading and spatial span scores by converting them into Z scores and averaging them to create a single WM span score. The analysis revealed a significant negative relationship between WM span and attachment preference, confirming that participants who had higher WM spans were more likely to prefer low attachment, and WM capacity modulated attachment preference across the three samples. While there were small differences in the relationship between reading and spatial WM tasks, these differences did not reach statistical significance, suggesting that the effect was somewhat stronger for the reading span task. This analysis also showed that L1 and L2 proficiency had no effect on the attachment preferences of the Spanish sample, and although not included as a prediction, neither L1 nor L2 proficiency showed association with attachment preferences among Mandarin speakers either.

## Internal validity

According to Morling, when making an association claim, internal validity is not relevant since there is no causality being claimed. That said, the authors of this study did take pains to ensure that they performed counterbalancing, avoided validity threats, and included covariates to ensure validity. For example, the researchers used the same paradigm from Swets et al. (2007, Study 1) in order to directly compare results and to reduce fatigue threats—they would have liked to have expanded the tasks further to incorporate new standards emerging in this area of study, however that would have made the experiment too long for the online environment.

## External validity

To determine external validity in an association claim, Morling asks us to consider how representative is the sample, and to what other situations might the association be generalized. Since this study was replicating and extending a previous study, one way they ensured external validity was to align tasks with the study being replicated (Swets et al., 2007). However, since they wanted to further their generalizability, they extended the previous research by looking at both high-low (L1 Spanish-L2 English) and low-low (L1 Mandarin-L2 English) combinations. However, potentially impacting that generalizability is the notion of heritage bilingualism. There's a difference in the ways in which the different language groups may have acquired their second language. Heritage speakers are early bilinguals whose native language is spoken at home and their second language becomes their primary language, the one learned and then used in school. This is common here in California, most notably in households that speak Spanish where the children go to schools that are taught only in English. Mandarin bilinguals are more likely to be late bilinguals, who learned English in early adulthood when using it in an academic setting like college, possibly raised in China before moving to the United States in mid- to late-adolescence. It would be good to find a heritage sample of Mandarin speakers to compare to in a future study to continue this exploration.

## Ethical considerations

Overall, this study appears to have been thorough and ethical. This study was clearly peer-reviewed, and they addressed reviewer's concerns in the journal article. The authors obtained approval by the ethics committee of the University of California, Davis, adhered to the principles of the Declaration of Helsinki, and appear not to have any conflicts of interest. Informed consent both for participation and for publication was obtained from all participants prior to beginning. Participants received course credit for participation. The study did not directly address how confidentiality would be maintained, how they might mitigate potential harm and provide support as needed, whether there was a debrief or follow-up, or how vulnerable populations might be addressed, however since this was a psycholinguistic study there wasn't much risk in any of these categories.

## Conclusion

Overall, this appears to be a strong study. The authors did a good job replicating and extending an existing study, making it stronger in the process. Their conclusions represented a reasonable interpretation based on the data and analysis. Of special note is that the study did not find what the authors expected in terms of Spanish sample—they thought that their English proficiency would impact their attachment preference, shifting them from a high attachment to a low attachment preference, however that was not the case. The authors thought this might be due to the impact of heritage bilingualism on the sample, which of course means there is room for further study here, but I thought that the fact that they stated their prediction upfront and then it didn't turn out to be back up by the data made it an even stronger study.

The limitations of this study are solid grounds for future research. As noted previously, the LHQ 3.0 is a self-report measure and therefore an objective measure could strengthen these results, however this measure was chosen as it has well-documented correlations with other established measures, supporting its validity. This study was limited to online during its timing during the pandemic—in the future an in-person study could enable them to do a different form of proficiency measure and further extend the tasks, however doing it online was a good call for this study and the choices they made prevented participant fatigue. Again, this seemed like a strong study—thoroughly thought out, well-executed, ethical, and solidly analyzed. I look forward to seeing how future research further extends these results.

## **References**

Cotter, B. T., & Ferreira, F. (2024). The relationship between working memory capacity, bilingualism, and ambiguous relative clause attachment. *Memory & Cognition*, *52*(7), 1530–1547. https://doi.org/10.3758/s13421-024-01561-4

Morling, B. (2021). *Research methods in psychology* (4th ed.). WW Norton.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, D.C.: U.S. Government Printing Office.