SAND2021-11681

SANDIA REPORT
SAND20XX-XXXX
Printed Click to enter a date

Sandia National Laboratories

# Physiological Characterization of Language Comprehension

Laura E. Matzen, Mallory C. Stites, Christina L. Ting, Kyra L. Wisniewski & Breannan C. Howell

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

**NOTICE:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

| | |
|---|---|
| Telephone: | (865) 576-8401 |
| Facsimile: | (865) 576-5728 |
| E-Mail: | reports@osti.gov |
| Online ordering: | http://www.osti.gov/scitech |

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

| | |
|---|---|
| Telephone: | (800) 553-6847 |
| Facsimile: | (703) 605-6900 |
| E-Mail: | orders@ntis.gov |
| Online order: | https://classic.ntis.gov/help/order-methods/ |

NNSA
National Nuclear Security Administration

# ABSTRACT

In this project, our goal was to develop methods that would allow us to make accurate predictions about individual differences in human cognition. Understanding such differences is important for maximizing human and human-system performance. There is a large body of research on individual differences in the academic literature. Unfortunately, it is often difficult to connect this literature to applied problems, where we must predict how specific people will perform or process information. In an effort to bridge this gap, we set out to answer the question: can we train a model to make predictions about which people understand which languages? We chose language processing as our domain of interest because of the well-characterized differences in neural processing that occur when people are presented with linguistic stimuli that they do or do not understand. Although our original plan to conduct several electroencephalography (EEG) studies was disrupted by the COVID-19 pandemic, we were able to collect data from one EEG study and a series of behavioral experiments in which data were collected online. The results of this project indicate that machine learning tools can make reasonably accurate predictions about an individual's proficiency in different languages, using EEG data or behavioral data alone.

# ACKNOWLEDGEMENTS

The authors would like to acknowledge Mike Trumbo and Alisa Rogers for assistance with data collection. We would also like to acknowledge our Academic Alliance collaborators in Dr. Kara Federmeier's lab at the University of Illinois at Urbana-Champaign. Special thanks to Dr. Federmeier and to Lin Khern Chia for their helpful input and productive collaboration on this project.

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

## CONTENTS

This page left blank

# ACRONYMS AND DEFINITIONS

Click here, then press delete to remove guidance.

**Required**. Spell out on first use in document and only include acronyms if used more than once. Only capitalize proper nouns.

| Abbreviation | Definition |
|---|---|
| AMT | Amazon Mechanical Turk |
| EEG | Electroencephalography |
| ERP | Event-related Potential |
| ML | Machine Learning |
| ms | milliseconds |

# 1. INTRODUCTION

Electroencephalography (EEG) records the electrical activity of the brain with millisecond-level resolution. It can be time-locked to events of interest, such as the onset of a stimulus in a person's environment, providing detailed information about how that stimulus was processed by the brain. These signals, called event-related potentials (ERPs), are well-characterized and are highly consistent across individuals. ERPs can reveal which languages a person understands, which could be useful in several applied contexts and for the general advancement of research on neurolinguistics and bilingualism. However, it is not clear whether an ERP-based language assessment method is feasible. The initial goal of this project was to assess the feasibility and limitations of such an approach.

However, the original goals of this project were severely disrupted by the COVID-19 pandemic. Just as our initial ERP study was ready to go, the pandemic shut down all in-person human subjects research. Our team pivoted to a new approach: collecting behavioral data online using Amazon Mechanical Turk. This required a completely new approach to the problem. EEG experiments do not require any overt responses from participants. We can simply analyze the participants' brain activity as they read or listen to linguistic stimuli. However, to collect behavioral data online, we needed to develop tasks where participants produced overt behavioral responses for every trial. Furthermore, we needed to develop language tasks that could be completed by participants even if they did not understand the language in question. Because of these constraints, a substantial proportion of our project was devoted to developing, testing, and modifying behavioral tasks for remote data collection. We identified some tasks that produced useful behavioral data and other tasks that did not. Each task and the results are outlined in the sections that follow.

After collecting multiple datasets using different behavioral tasks, we applied machine learning (ML) techniques to the datasets in order to develop models of performance for monolingual and bilingual individuals. These models were trained using response time data from our online tasks, then used to predict which participants fell into each category. For some of our tasks, the models were able to make fairly good predictions, matching the participants' self-reported proficiency 70% of the time or more. The modeling approach also allowed us to examine different features that contributed to the model's predictions. This led to some interesting insights into items that were especially good at differentiating between the two groups as well as cases where the model struggled. A particularly interesting finding was that the model often struggled with individuals who were bilingual but had learned their second language later in life. This suggests that even simple response time measures may be sensitive to effects of the age of language acquisition.

For other datasets, the ML approach did not work as well, indicating that not all of our tasks were successful in producing stable results that reflect differences in language comprehension. By comparing different tasks, we were able to assess how well the results of each task reflected participants' self-reported language proficiency. This provides useful information for future task development for research on bilingualism, language proficiency, or other individual differences.

During the summer of 2021, our team was able to collect data for one of the EEG experiments that was originally planned for this project. The results of this experiment suggest that our original hypothesis, that ERPs can be used to characterize individual differences in language processing, was supported. Our findings show that further research in this area is warranted and could lead to useful advances in both cognitive science research and applications.

# 2. WORD LENGTH JUDGMENT EXPERIMENT[1]

## 2.1. Background

Studies of bilingual language processing have raised interesting questions about the nature of linguistic and semantic representations in semantic memory. Many open questions remain regarding the organization of multiple languages within the processing system, particularly the extent to which two languages share underlying conceptual representations and automatically activate one another during processing. Much of this research has relied on the use of priming paradigms to probe the size and nature of cross-language priming effects, as a way to understand whether the bilingual processing system shares representations across languages, or whether concepts may be represented separately. Two competing models of bilingual language comprehension, the Bilingual Interactive Activation (BIA+) model (Dijkstra & van Heuven, 2002) and the Revised Hierarchical Model (RHM; Kroll & Stewart, 1994; Kroll, van Hell, Tokowicz, & Green, 2010), have been proposed to help account for varying priming effects observed across studies.

One of the most common tasks used to probe the nature of bilingual language processing is cross-language translation priming. Masked repetition priming effects are well-established within a speaker's native language (Forster & Davis, 1984), particularly in the lexical decision task. By comparing the size of within- and across-language translation priming, researchers can begin to understand how effectively words in one language facilitate the same concept in their second language. Cross-language non-cognate translation priming effects have also been observed in cases where priming from a word in one language facilitates responses to that word's translation in another language (e.g., Grainger & Frenck-Mestre, 1998). Translation priming effects tend to be bigger under certain circumstances: for example, in more proficient bilinguals, with longer prime durations, and with priming from L1 primes to L2 targets (rather than L2 primes to L1 targets), (see Schoonbaert, Duyck, Brysbaert, & Hartsuiker, 2009 for review).

Research in this field has traditionally relied on recruiting groups of individuals with known language backgrounds and testing how priming effects manifest in these pre-established language proficiency groups. However, this existing paradigm is not without challenges. An individual's language background is hard to effectively quantify, and individuals can vary widely in their second language proficiency even within relatively well-matched groups (for review, see van Hell & Tanner, 2012). Because the size or presence of priming effects depends heavily on correctly characterizing participants' language background, it seems that efforts to establish a more individualized approach to data analysis could help the field identify more consistent findings, in turn advancing our understanding of the bilingual language processing system.

The present study seeks to establish a novel approach to bilingual language comprehension research that capitalizes on individual differences rather than averaging over them. We are interested in trying to characterize an individual's language background, without knowing it in advance, based on their behavioral responses in a cross-language priming task. Specifically, we utilize supervised machine learning techniques to identify patterns in response time data that may differentiate individuals who are proficient in the target language from those who are not. This approach represents a departure from traditional paradigms and leverages cross-disciplinary data analysis techniques to provide a potential new avenue for the study of bilingual language processing.

[1] This section was published in the proceedings of the Cognitive Science Society Annual Meeting 2021.

In order to collect behavioral responses to words in a language an individual may not know, we employed a word-length judgment task rather than lexical decision or semantic categorization. This task has been successfully used to elicit N400 priming in a bilingual population in which L1 and L2 words were intermixed (Martin, Dering, Thomas, & Thierry, 2009), indicating that the task could still allow for contact with the word's semantics. Additionally, as in Martin et al. (2009), we intermixed trials from the two languages rather than using the more typical blocked design. This choice was made to make it less predictable at the trial level whether the upcoming word would be in a language the participant knew, thus further encouraging participants to access each word's semantics. We predicted that we would see within-language repetition priming effects for the languages in which the participant was proficient. We also predicted that proficient bilinguals would show translation priming effects, whereas participants who were not proficient in the second language would not show these effects. Furthermore, exploratory machine learning analyses will allow us to test whether other aspects of the behavioral data could reliably predict an individual's language proficiency.

## 2.2. Methods

This study and all other studies that were a part of this project were reviewed and approved by the Human Studies Board at Sandia National Laboratories. A total of 95 participants were recruited via Amazon Mechanical Turk (AMT). To qualify for the task, the participants had to have an approval rate >95% for prior tasks completed on AMT. A subset of 40 participants also met AMT's criteria for fluency in Spanish. Participants were paid $3-4 for their time.

### *2.2.1. Materials*

The materials consisted of 30 Spanish nouns and their translations in English. The words were selected so that there were no special characters (accents, etc.) and no cognates or false cognates. We took care to select Spanish words that monolingual English speakers would be unlikely to encounter in their daily lives. Using information from the CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012), the word lists were matched on length, frequency, and orthographic neighborhood size, as shown in Table 1. The orthographic neighborhood size across languages was minimized.

**Table 1: Matched Properties of the Two Word Lists**

| | English | Spanish |
|---:|:---:|:---:|
| Avg. Length | 5.37 | 5.47 |
| Avg. Frequency | 106.93 | 99.04 |
| Avg. English Orthographic Neighborhood Size | 6.77 | 0.77 |
| Avg. Spanish Orthographic Neighborhood Size | 2.27 | 5.43 |

The words were paired in eight types of pairings: English-English repetitions, Spanish-Spanish repetitions, English-Spanish translations, Spanish-English translations, English-English unrelated, Spanish-Spanish unrelated, English-Spanish unrelated, and Spanish-English unrelated. There were a total of 240 pairs, with each word appearing in every possible pair type, four times as a prime and four times as a target. The word pairs were divided into four blocks of 60 pairs each. Each target word appeared twice in each block, once as part of a related pair and once as part of an unrelated pair. The pairs were placed in the pseudorandom order so that the two pairs that contained the same target word appeared in different halves of the block. The pseudorandom order was constrained so that there were never more than four translation/repetition or unrelated pairs in a row, and never more than two pairs of the same type (e.g., Spanish-English translation) in a row.

### *2.2.2. Procedure*

After reading and acknowledging the consent form, participants completed a short language proficiency questionnaire with questions that were similar to those in the Language Experience and Proficiency Questionnaire (Marian, Blumenfeld & Kaushanskaya, 2007). They were asked to list up to four languages that they know, first in order of dominance and then in order of acquisition. They were asked what percentage of the time they are currently exposed to English and Spanish, and how much total time they have spent living or traveling in countries where Spanish or English is the dominant language. Finally, they were asked to rate their level of proficiency in English and Spanish on an 11-point scale ranging from "None" to "Perfect," the age at which they began to acquire each language (infant, child, teen, adult, or never), and which factors contributed to them learning that language. The response options included interacting with family, interacting with friends, formal language classes, reading, language tapes/learning apps/self-instruction, watching TV or movies, listening to the radio, and travel.

After completing the questionnaire, participants were shown the task instructions and an example. They were told that they would see words in English and Spanish, and that the words would sometimes be repeated or followed by the same word in the other language. They were told to press the "B" key on the keyboard if the word had 5 letters or fewer and the "N" key if the word had 6 letters or more. They were instructed to respond as quickly as possible without making too many mistakes. Finally, the participants were told that there were four blocks of words with breaks in between, and that each block would take about two minutes to complete. When they were ready to begin, they clicked on a button labeled "Start Experiment." The first six words that the participants saw were practice words and were not included in the analysis. The participants responded to every item, whether it was a prime or a target.

## 2.3. Behavioral Results

A total of 13 participants were excluded from the analysis, either because they did not complete the entire task, they did not provide consistent responses to the questionnaire, or because their pattern of responses indicated that they were responding randomly rather than following the task instructions. Of the remaining 82 participants, 40 were from the group that met AMT's criteria for fluency in Spanish and 42 were from the group with no specific language qualification requirements.

In the group that met AMT's criteria for fluency in Spanish, one participant rated his/her proficiency in reading Spanish at 7 ("Good"), and all of the other participants rated their proficiency at 8 ("Very Good") or higher on the 0-10 scale. Thirty-three of the participants in this group reported that Spanish was their dominant language and the first language they acquired. Three participants reported that English was their dominant language and the first language they acquired. Two participants reported that Spanish was the first language they acquired, but English was their dominant language. Two participants reported that English was the first language they acquired, but Spanish was their dominant language. All of the participants in this group reported that they had lived for at least one year in an area where Spanish is the predominant language (range 1-57 years, mean = 28.8 years). They had spent an average of nine years living in areas where English was the predominant language (range = 0-54 years). Thirty-three of the participants reported that they had spent more time living in predominantly Spanish-speaking areas than in predominantly English-speaking areas, and 17 reported that they had never lived in an area where English was the predominant language.

In the group of participants that was recruited without the use of AMT's Spanish fluency qualification, all of the participants reported that English was their dominant language, and all but one of the participants reported that English was the first language they acquired (one person reported that their first language was Mandarin). There were 21 participants who reported that they did not know any Spanish at all. Another 15 participants reported that they had learned some Spanish as a teen or adult, primarily through formal language classes or self-instruction, but they rated their Spanish proficiency at 3 ("Fair") or below. Three participants reported that they began learning Spanish as teenagers and gave themselves intermediate fluency ratings (5-7). Finally, three participants rated their Spanish proficiency as 8 or higher. One of these participants reported that they started learning Spanish in infancy, one in childhood, and one as a teen. The participants reported that they had spent an average of 37.7 years living in predominately English-speaking areas (range 25-70 years) and an average of 3 years living or traveling in predominantly Spanish-speaking areas (range 0-23 years).

For our analyses, we grouped all of the participants who rated their Spanish proficiency as 8 or higher into the "proficient" group, regardless of whether or not they had AMT's qualification for Spanish proficiency. There were a total of 42 participants in this group, 39 from the batch that required the AMT Spanish qualification and three from the batch that did not. All of the participants who rated their Spanish proficiency at 7 or lower (40 participants) were placed in the "non-proficient" group. One of these participants was from the batch that required the AMT Spanish qualification and the 39 were from the batch that did not.

We began with a traditional analysis of the priming effects for each experimental condition. The participants' average response times were calculated for each condition. Only correct trials were included in the analysis. Trials with response times (RTs) of less than 200 milliseconds were excluded, as were trials with RTs that were more than three standard deviations higher than that participant's mean response time (unless those trials had RTs that were less than 6 seconds). A total of 111 trials out of 19,680 were excluded due to having unusually short or long response times. For each participant, the priming effect for each condition (English-English, Spanish-Spanish, English-Spanish, Spanish-English) was calculated by subtracting the average RT for the targets in the repetition or translation pairs from the average RT for the targets in the unrelated pairs. Figure 1 shows the average size of the priming effects across participants.

A 2 (Spanish Proficiency) x 4 (Priming Condition) ANOVA showed that there was a significant effect of proficiency group ($F(1,240) = 16.77, p < .001$), a significant effect of condition ($F(3,240) = 90.04, p < .001$), and a significant interaction between the two ($F(3, 240) = 5.11, p < .01$). The participants in the proficient Spanish group had a significantly larger priming effect than the other group for both the English-English ($t(67) = 3.50, p < .001$) and the Spanish-Spanish condition ($t(67) = 3.59, p < .001$). For the two cross-language conditions, neither group showed a priming effect and the two groups did not differ significantly from one another (all *t*s < 1.12, all *p*s > .13).

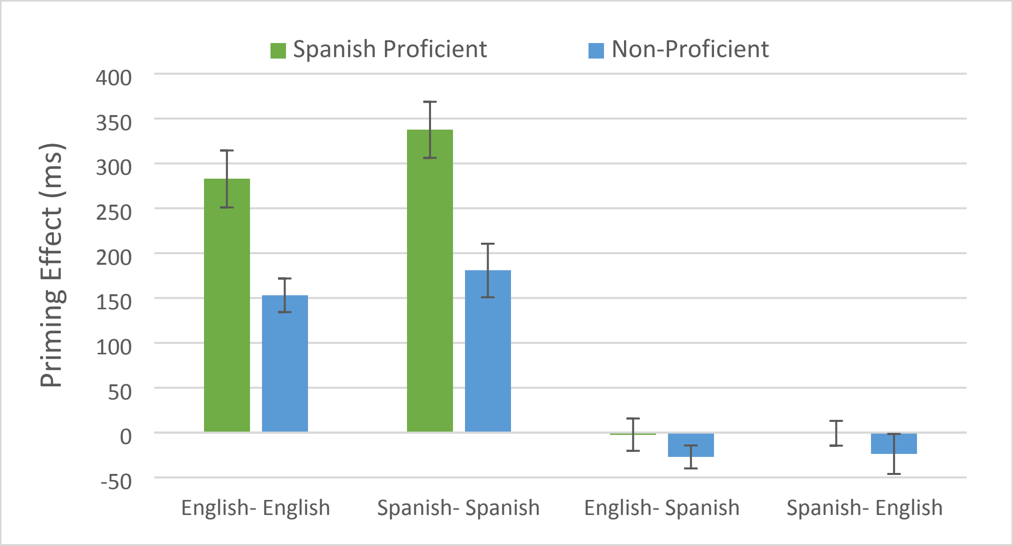

**Figure 1. The average magnitude of the priming effects. Error bars show the standard error of the mean**

## 2.4. A Model of Bilingual Language Proficiency

The priming effects (PEs) showed that there was a significant difference between participant groups in the English-English and Spanish-Spanish priming conditions. A potential application of this result may be to learn a function that maps the priming effects of known participants to their corresponding proficiency labels so that we can use the priming effects from new participants to predict their proficiency.

More generally, *classification* is a standard supervised machine learning (ML) task that follows a *train* and *predict* paradigm. During the training phase, labeled data is used to build a model (i.e., a learned function) that maps an input (typically numerical feature vectors) to an output (labels). During the predict phase, the model is used to infer the labels of new data (James, Witten, Hastie, & Tibshirani, 2013).

An advantage of this approach is that ML algorithms can usually handle very high dimensional data (e.g., the individual PEs or RTs) compared with standard statistical analyses of behavior, which look at averages (e.g., average PE or RT for a particular condition). A disadvantage of this approach is that ML algorithms are often considered "black boxes", providing very little interpretability as to how the model arrives at its prediction.

A linear Support Vector Machine (SVM), on the other hand, is a simple but successful ML algorithm that yields insights as to how the individual features (e.g., PEs and RTs) contribute to the predicted output (Boser, Guyon, & Vapnik, 1992; Cristianini & Shawe-Taylor, 2000). In its simplest form, the objective of an SVM is to find a hyperplane that separates the labeled data into the two distinct classes (extensions for multiclass problems exist), while also maximizing the distance between the hyperplane and the nearest point from either group (hard-margin). The coordinates of the vector orthogonal to the hyperplane form the weights (coefficients) of the model. From the weights, it is possible to do two things. First, we can determine feature importance according to the relative magnitude of the weights. Second, new data items can be labeled depending on which side of the hyperplane they fall (computed by taking the dot product with the orthogonal vector).

For our application, we use the Linear Support Vector Classification (LinearSVC) class available in Python's Scikit-learn 0.23.1 with default parameters. Scikit-learn 0.23.1 is used throughout our ML workflow for data preprocessing, feature engineering, and model validation  (Pedregosa et al., 2011).

### 2.4.1. Data Preprocessing

Using the same criteria as in the prior section, participants were assigned proficiency labels based on their survey responses. Specifically, 42 participants were labeled as "proficient" in Spanish and 40 participants were labeled as "non-proficient" (English proficiency is assumed).

Each participant was associated with a list of 240 RTs for each of the 240 target words in the experiment. Across all participants, the mean RT for the target words was 825 ms and the standard deviation was 231 ms. Target words with a mean RT that was more than three standard deviations above this mean were removed from the dataset for all participants. Only one target word was excluded based on this criterion, leaving us with 239 RTs for each participant. Then, to account for different baseline RTs for different participants, each participant's RTs were normalized from 0 to 1.

We note that this approach for preprocessing the data for input into the SVM differs from the approach for cleaning the data for the behavioral analysis. In the behavioral analysis, each participant's data is cleaned by removing individual trials with incorrect responses and/or unusually short/long responses. Thus, each participant is left with a *different* set of RTs and PEs after cleaning the data. However, for input into the SVM, each participant must be represented by the *same* set of features, necessitating a different approach to removing anomalous data.

### 2.4.2. Feature Engineering and Selection

From the 239 normalized RTs, we construct feature vectors that are used as input into the SVM as follows. The first feature set simply represents the 239 normalized RTs. The second feature set represents the PEs. Each English target word appears in two PEs (English-English and Spanish-English); similarly, each Spanish target word appears in two PEs (Spanish-Spanish and English-Spanish). Therefore, for the 60 target words in this study, we have 120 PEs. Because one target word was excluded, we are left with 119 PEs.

Given a set of features, a standard next step in a machine learning workflow is to perform some type of feature selection technique to reduce the number of features, i.e. reduce the dimensionality. Reducing the number of features, particularly when the number of features exceeds the number of samples, can improve the accuracy of the model.

Univariate feature selection is one of the simplest techniques to reduce the number of features and works by selecting the best set of features based on univariate statistical tests such as a chi-squared test or an ANOVA. We will use an ANOVA to compute the p-value between the label and features to select the m best features according to the lowest p-values.

### 2.4.3. Model Validation

In a deployed setting, we would apply our SVM model that has been trained on the 82 participants of known proficiency to make predictions on new participants of unknown proficiency. However, without validating the model first, it is not possible to know how good the new predictions are. Therefore, a cross-validation test is usually performed first, in which part of the labeled data is withheld during training and used to test (validate) the performance of the model during prediction. Many methods exist to split the data into train/test sets. Perhaps most common is the *k*-fold cross validator, which splits the data into *k* consecutive folds. Each fold is then used once as the test (validation) set, while the remaining *k* – 1 sets form the training set. We use *k = 5* and perform 10 runs of each of the cross-validation experiments.

Finally, the model (i.e. *m* best features) with the highest mean *balanced accuracy score* is selected. The balanced accuracy is defined as the average accuracy obtained on each class (non-proficient, proficient) and is used in place of accuracy when there is a class imbalance (Brodersen, Ong, Stephan & Buhmann, 2010).

### 2.4.4. Results Using Priming Effect Size

We begin with prediction results using the PEs as the features. **Error! Reference source not found.** shows the mean and standard deviation (SD) of the balanced accuracy as a function of the *m* best PE features used to train the SVM. We achieve the highest accuracy of 0.68 (SD = 0.11) with *m* = 98 features. For comparison, an accuracy of 0.62 (SD = 0.10) is achieved using the average PEs as features.

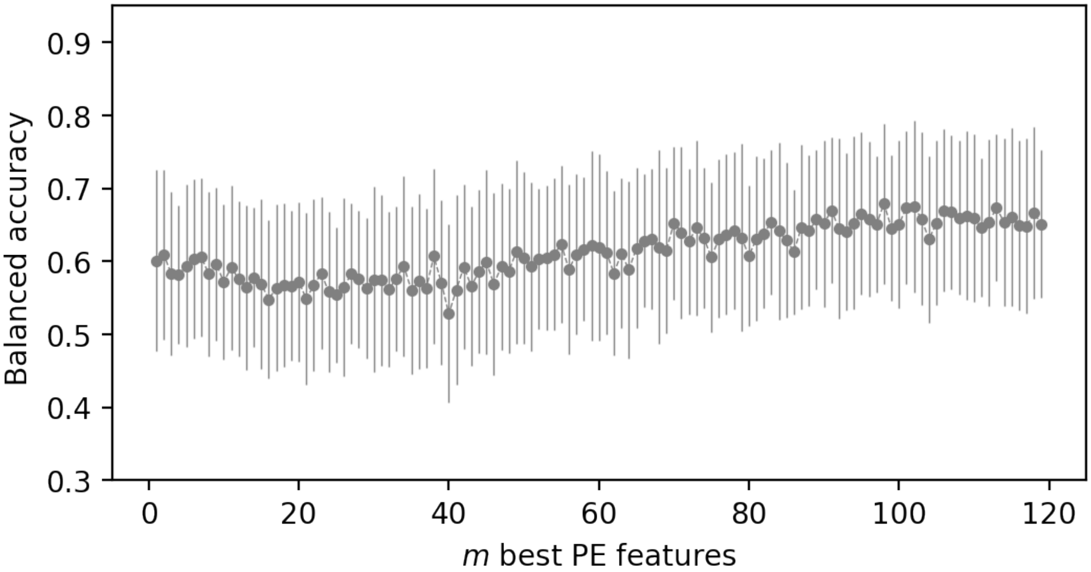

**Figure 2. Mean and standard deviation of the balanced accuracy as a function of the *m* best priming effect (PE) features used to train the SVM. Best performance (mean accuracy = 0.68) is achieved at *m* = 98.**

In Table 2, we also show the mean and the standard deviation (parentheses) of the confusion matrix for the best-performing model using *m* = 98 features. The confusion matrix shows the class-level prediction accuracy. From these results, we can see that the model predicts the non-proficient participants with slightly higher class accuracy than the proficient participants.

We would also like to understand how the different PEs contributed to the proficiency prediction of the SVM. Figure 3 plots the mean values for the two metrics for significance for each of the 119 PE features. The SVM weights correspond to weights *after* feature selection. If a feature is not chosen it is given a weight of 0. In general, the features with the highest SVM weights also have small *p*-values. This result supports the intuition that features with lower *p*-values should also contribute more predictive power (higher weights) to the SVM model. Interestingly, three of the top four most predictive features (by either metric) correspond to the words CUELLO, LLUIVA, and PILLOW.

All three of these words are six letters long and contain the digraph 'll,' which was considered to be a distinct letter in the Spanish alphabet prior to 2010 (Real Academia Española, 2010). In our word length task, participants were asked to press one button for words that were five letters or shorter and another for words that were six letters or longer. Given this task and the relatively recent removal of 'll' from the Spanish alphabet, these three words may have been tricky for the proficient Spanish speakers. It is notable that the model identified these three stimuli as the ones that were most effective for differentiating between the two groups of participants.

**Table 2. Mean and standard deviation (parentheses) of the confusion matrix for the best performing PE model.**

| | | Predicted Group | |
|---|---|---|---|
| | | Spanish Proficient | Non-proficient |
| Actual Group | Spanish Proficient | 0.61<br>(0.16) | 0.39<br>(0.16) |
| | Non-proficient | 0.26<br>(0.16) | 0.74<br>(0.16) |

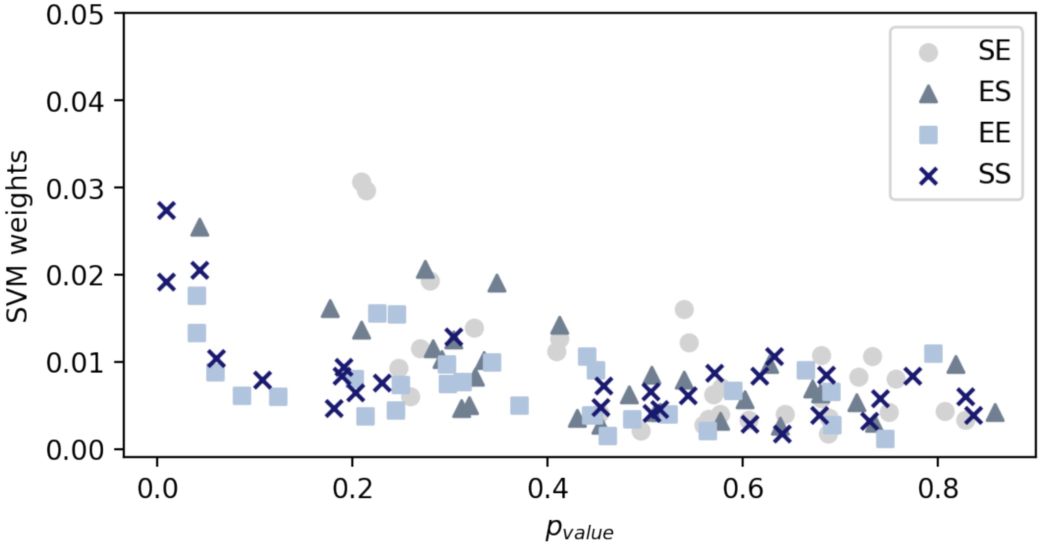

**Figure 3. Priming effect (PE) feature significance. Features with low *p*-values (significant according to the univariate statistical test) and high coefficients (significant according to the model) are the most predictive.**

### *2.4.5. Results Using Response Times (RTs)*

Next, we repeat our analysis using response times (RTs) as features for the SVM. Figure 4 and Table 3 show the prediction performance the SVM classifier using RTs as features. Overall, we achieve better performance using RTs, compared with using PEs, as features. We achieve the highest balanced accuracy of 0.75 (SD = 0.09) with *m* = 175 features. For comparison, a balanced accuracy of 0.66 (SD = 0.11) is achieved using the average RTs as features.

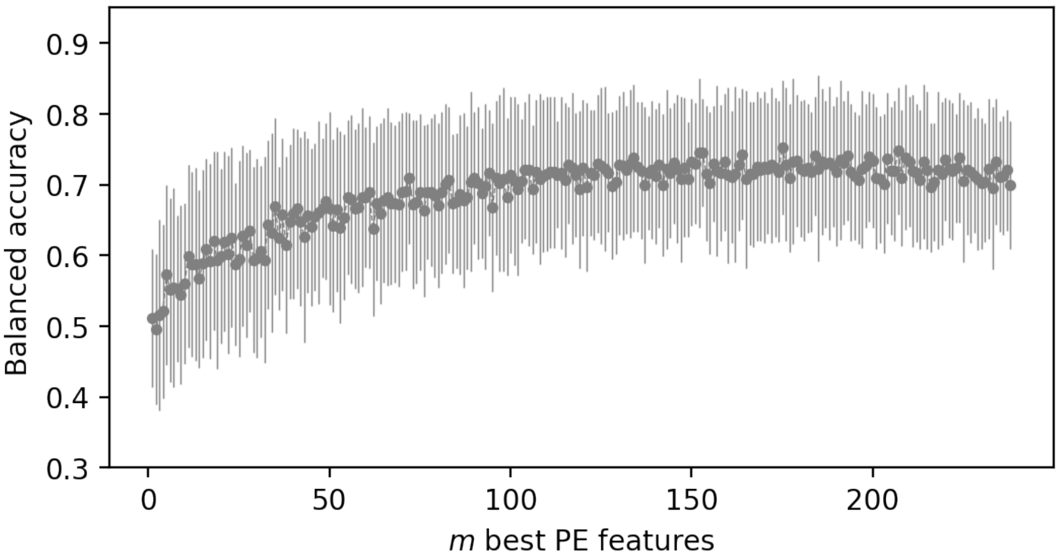

**Figure 4. Mean and standard deviation of the balanced accuracy as a function of the m best response time (RT) features used to train the SVM. Best performance (mean accuracy = 0.75) is achieved at m = 175.**

**Table 3. Mean and standard deviation (parentheses) of the confusion matrix for the best performing model.**

| | | Predicted Group | |
|---|---|---|---|
| | | Spanish Proficient | Non-proficient |
| Actual Group | Spanish Proficient | 0.74<br>(0.15) | 0.26<br>(0.15) |
| | Non-proficient | 0.23<br>(0.13) | 0.77<br>(0.13) |

As with the PEs, we would like to understand how the individual features contribute to the ability of the SVM to predict participant proficiency. Figure 5. Response time (RT) feature significance. Features with low p-values (significant according to the univariate statistical test) and high coefficients (significant according to the model) are the most predictive.

5 plots the *p*-value and the mean SVM weight for each of the 239 RT features. Once again, in general, RT features with higher SVM weights have smaller *p*-values, indicating that features with lower *p*-values tend to contribute more predictive power (higher weights) to the SVM model.

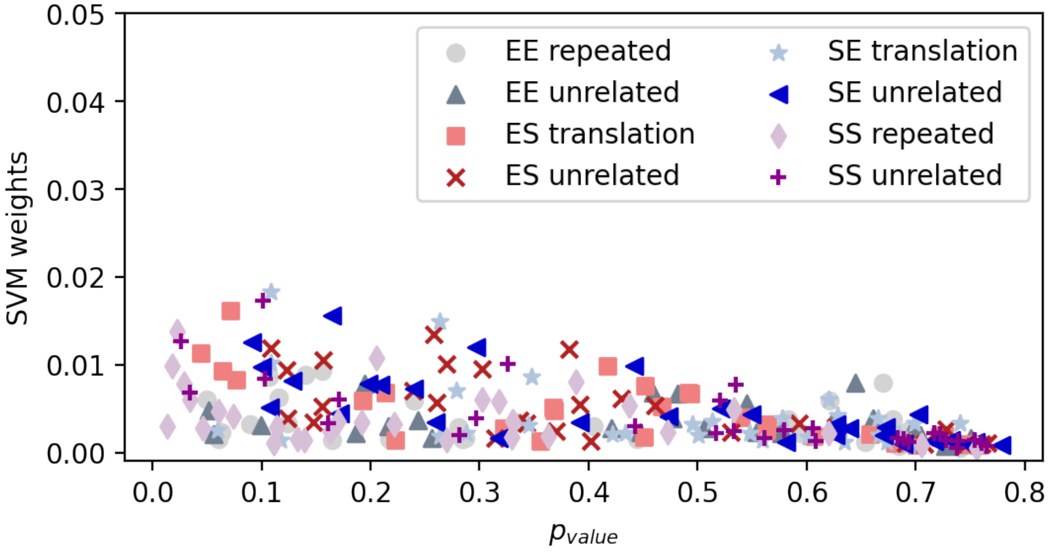

**Figure 5. Response time (RT) feature significance. Features with low p-values (significant according to the univariate statistical test) and high coefficients (significant according to the model) are the most predictive.**

We also examined which participants were misclassified in the highest-performing version of the model. Interestingly, there were six proficient Spanish speakers who reported that they started learning English before learning Spanish and that English was their dominant language. Four of those participants were consistently misclassified by the model, which placed them in the non-proficient group 90-100% of the time. Another participant in this group was misclassified 30% of the time. Only one participant in this category was always classified as being proficient in Spanish, and that was also the only participant who reported that they learned both English and Spanish beginning in infancy. The others in this subset began learning Spanish later in childhood or as teenagers. Although some of these participants may have simply overstated their Spanish proficiency, this pattern suggests that age of acquisition could be a key factor in the RT effects that are identified by the model.

## 2.5. Discussion

This study employed a repetition and translation priming paradigm to test the efficacy of using machine learning techniques to characterize an individual's language proficiency based on priming data. Our analyses showed within-language repetition effects for both languages, with priming effects that were larger for proficient Spanish speakers. However, we observed no priming effects for translations, suggesting that our effects were driven by the wordform and/or response priming, rather than semantic priming. On the surface, these findings may provide weak support for the RHM model (Kroll & Stewart, 1994) because we do not see facilitation between translation equivalents even for people who are proficient in both languages. However, our experimental paradigm and non-semantic task may have encouraged shallow processing. Unlike a classic priming paradigm, where participants see a prime and then respond to a subsequent target, the participants in our task responded to every word with no differentiation between primes and targets. Due to this design and the intermixing of within-language and cross-language pairs, the participants may have been less likely to make predictions about which word would come next, which could reduce the effect of semantic priming. In future research, we plan to test blocked designs where all of the targets in each block are in the same language and a more traditional priming paradigm in which participants passively read the primes and respond only to the targets. We predict that those changes to the experiment structure will produce larger semantic priming effects for proficient bilingual participants reading translated pairs.

Our machine learning analyses showed that a model trained on reaction time data and priming data can predict whether an individual participant is proficient in Spanish with high accuracy. Interestingly, for this dataset, predictions based on priming effects were slightly less successful than predictions based on the RTs alone (68% versus 75% prediction accuracy). Even though the experimental task may have encouraged shallow processing, the participants who acquired Spanish beginning in infancy displayed patterns of response times that differentiated them from the other participants. The model also revealed specific words that were more predictive of proficiency than others, indicating that this approach could also be fruitful for item analyses.

This study has several limitations. Most importantly, we based the proficiency labels on the self-reports on anonymous online participants. The majority of the participants (39 of 42) who reported high proficiency in Spanish also had a Spanish fluency qualification from Amazon Mechanical Turk, which provides some external verification of their proficiency. However, it is not clear what criteria are used to assign that qualification. In future research, it would be useful to assess the model's performance against measures of language proficiency that are more objective than self-reporting.

The word length judgment task that we used also has limitations. We were constrained to using a task that all participants could complete whether they understood Spanish or not. In future work, we aim to develop new tasks that can be completed without knowledge of the target language but that encourage semantic processing.

Overall, this study demonstrates that machine learning techniques can support a more individualized approach to data analysis in studies of bilingualism or other individual differences. Rather than simply averaging data from all of the participants within each group and comparing the two groups, the ML approach allows us to develop a predictive model to classify participants based on their language proficiency, as instantiated in the data they produced. This can be used to identify groups of participants with different proficiency levels, rather than assigning participants to groups in advance, or to explore differences among participants with similar levels of proficiency. Finally, machine learning can be used to identify the specific stimuli that are most predictive of participant proficiency. All of these factors enable new approaches to the study of bilingualism.

# 3. WORD LENGTH JUDGMENT TASK WITH CLASSIC SEMANTIC PRIMING DESIGN

In this experiment, we modified our original word length task in an effort to achieve larger cross-language semantic priming effects. We altered the task so that it was more similar to traditional semantic priming experiments, in which participants passively read a prime word and then respond to a related or unrelated target word. We predicted that when participants viewed the primes, they might make predictions about what word was coming next, potentially leading to larger semantic priming effects.

## 3.1. Methods

### *3.1.1. Participants*

Twenty-one people participated in this study on Amazon Mechanical Turk. The task was posted in two different batches, one which required participants to have Mechanical Turk's Spanish fluency qualification and at least a 95% approval rate for prior tasks completed on Mechanical Turk. To participate in the other batch, the participants were required to be located in the United States and to have completed at least 1000 prior HITs on Mechanical Turk with an approval rate of at least 95%. The participants were paid $3.50-$4 for their time.

Two of the participants were excluded from the analysis because they responded randomly rather than following the task instructions. The final set of participants included nine from the batch that required Mechanical Turk's Spanish fluency qualification and ten from the batch that did not.

### *3.1.2. Materials and Procedure*

The materials used in this experiment were identical to those used in the work length task. The key difference was that the participants only responded to the target words instead of responding to both words in the pair. The prime word appeared on the screen for one second in lowercase black font. Then the target word appeared in uppercase red font. The participants were instructed to press the "B" key on the keyboard if the target word was 5 letters or fewer and the "N" key if the target word was six letters or longer.

As in the prior study, the participants also completed a short survey about their language background before beginning the word length judgement task.

## 3.2. Results

In the group that met AMT's criteria for fluency in Spanish, one participant rated his/her proficiency in reading Spanish as "good" and all of the other participants rated their proficiency at "very good" or higher (8+ when converted to a 0-10 scale. All of the participants in this group also rated their English proficiency as being "good" or better. Six of the participants in this group reported that Spanish was their dominant language and the first language they acquired. Two participants reported that English was their dominant language and the first language they acquired, and one participant reported that Romanian was their dominant/first language. Six of the participants had lived for at least 15 years in an area where Spanish is the predominant language (range 15-47 years, mean = 19.7 years). Seven participants had lived in an area where English is the predominant language (range 3-34 years, mean = 14.78 years).

In the group of participants that was recruited without the use of AMT's Spanish fluency qualification, all of the participants reported that English was their dominant language, and all but one of the participants reported that English was the first language they acquired (one person reported that their first language was Mandarin). There were 8 participants who reported that they did not know any Spanish at all and two participants who reported minimal knowledge of Spanish, rating their proficiency as "low" or "very low" (1 or 2 on the 0-10 scale). All of the participants in this group reported having lived in a predominantly English speaking country for at least 10 years (range = 10-58, mean = 43.6 years). None of the participants had lived in a predominantly Spanish speaking country.

A 2 (Spanish Proficiency) x 4 (Priming Condition) ANOVA showed that there were no significant effects of proficiency group or condition. The participants in the non-proficient group showed priming effects that were near zero in all conditions. The participants in the Spanish proficient group showed negative priming effects in the same language conditions and modest positive priming effects in the cross-language conditions. The results are shown in Figure 6.

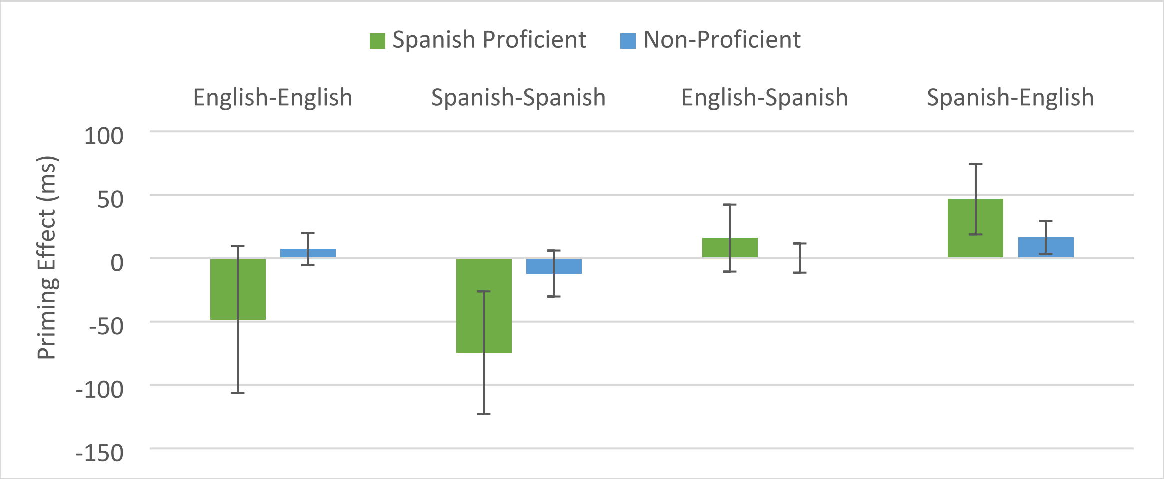

**Figure 6. The average magnitude of the priming effects. Error bars show the standard error of the mean**

Looking at the individual participants, there was only one Spanish-proficient participant who showed a positive priming effect in all four conditions, as would be expected for a bilingual Spanish-English speaker if our task effectively induced semantic priming. In the non-proficient group, where we would expect to see a priming effect in the English-English repetition condition, only six of the ten participants showed a positive priming effect.

### 3.3. Discussion

The results indicate that this task did not produce a reliable priming effect for either the proficient Spanish speakers or the monolingual English speakers. We predicted that using a more traditional priming task design where the participants passively viewed the primes and responded only to the targets would increase semantic processing of the words and lead to a larger priming effect than we observed in our initial study. However, this prediction was not supported. It seems that this task design did not encourage semantic processing of the words, and may have led the participants to ignore the prime words altogether, producing little to no semantic priming.

# 4. PICTURE PRIMING TASK

In another attempt to elicit semantic priming in a multilingual task where not all participants understood the linguistic stimuli, we developed a picture priming task. In this task, participants saw a prime word in English or Spanish followed by a picture. They were asked to press one key if the picture showed a living thing (animal, insect, etc.) and another if the picture showed a nonliving thing (furniture, household items, etc.).

## 4.1. Methods

### *4.1.1. Participants*

A total of 62 participants completed the task on Amazon Mechanical Turk (AMT). Twenty-two participants completed a version of the task that required AMT's Spanish fluency qualification. Forty participants completed a version of the task that did not require this qualification. Some of the participants in this group may have had the qualification, but it was not required for participation. All participants were required to have an approval rate greater than 95% for prior tasks completed on AMT. The participants were paid $2 for their time.

### *4.1.2. Materials*

The pictures used in this experiment were drawn from the MultiPic set of drawings, which have been normed in multiple languages, including English and Spanish (Duñabeitia, Crepaldi, Meyer, New, Plaitsikas, Smolka & Brysbaert (2018). Images of 20 living and 20 nonliving things were selected for use in this study. All of the images selected were consistently given the same names by participants in the norming study, both in English and Spanish. Each image was paired with its name in English and its name in Spanish, as well as an unrelated word from the opposite category (living or nonliving) in both English and Spanish. The result was a list of 160 word-picture pairs.

Within the stimulus list, each prime word appeared twice, once with the picture that matched that word and once with an unrelated picture from the opposite category. Each picture appeared a total of four times, preceded by the corresponding word in English, the corresponding word in Spanish, an opposite-category word in English, and an opposite-category word in Spanish. The word-picture pairs were placed in a pseudorandom order such that no more than two items from the same condition (match or mismatch) appeared in a row, nor more than four items with the same answer (living or nonliving) appeared in a row, and no more than four items with prime words in the same language appeared in a row.

### *4.1.3. Procedure*

At the beginning of the experiment, the participants completed the same language history questionnaire that was used in our prior experiments. Then they read the task instructions. They were told that on each trial they would see a word in English or Spanish that would stay on the screen for 1 second and that the word would be followed by a picture. The word and picture would match about half of the time, and their job was to press the "B" key on the keyboard if the picture showed a living thing and the "N" key if it showed a nonliving thing. They were told to try to respond as quickly as possible without making too many mistakes, and were asked to keep one finger resting on each key to facilitate speedy responses.

After reading the instructions, participants proceeded to the task. The first three trials were practice trials and were followed immediately by the 160 real trials. There was a self-paced break halfway through the real trials.

Each trial began with a blank screen that was presented for 150 ms, followed by the prime word, which was presented in lowercase black font. The word was presented for 1 second and was followed by a blank screen for 150 ms. Then the picture was presented, scaled to be 300 pixels wide. The picture remained on the screen until the participant pressed the “B” or “N” key, at which point the next trial began.

The words were presented in 20-point Arial font and appeared slightly above where the target pictures appeared on the screen. An example of one trial is shown in Figure 7. A reminder of which key corresponded to the “living” and “nonliving” response appeared above the stimuli at all times during the experiment in grey 12-point Arial font

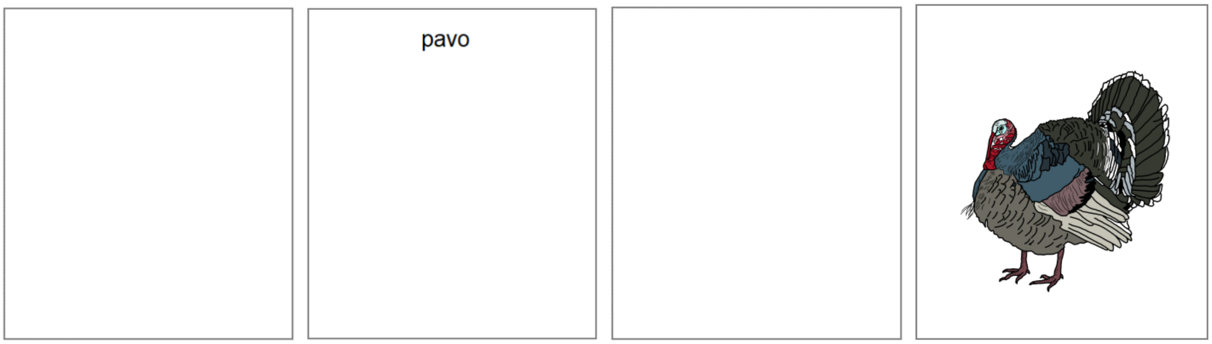

**Figure 7. An example trial. The blank screens appeared for 150 ms each, the prime word appeared for 1 second, and the picture remained on the screen until the participant responded.**

## 4.2. Results

The participants who completed the version of the task that required AMT’s Spanish fluency qualification all rated their proficiency in reading Spanish as being “very good” or higher. The participants in this group also reported having similarly high levels of proficiency in reading English. These participants reported that they had lived in predominantly Spanish-speaking countries for an average of 30.9 years (range = 5-55) and in predominantly English-speaking countries for an average of 6.2 years (range = 0-35).

For the group who completed the task that did not require AMT’s Spanish fluency qualification, all but two reported that their proficiency in reading English was “excellent” or “perfect.” The two participants who reported low proficiency in English were excluded from further analysis.

The participants in this group reported that they had spent an average of 28.2 years living in predominantly English-speaking countries (range = 0-77) and an average of 4.6 years living in predominantly Spanish-speaking countries (range = 0-49). There were nine participants who reported that they had never learned any Spanish at all. Another 17 participants rated their Spanish proficiency as being “fair” or lower. Seven participants gave themselves intermediate proficiency ratings (“slightly less than adequate,” “adequate,” “slightly more than adequate” or “good”). Finally, five of the participants rated their Spanish proficiency as being “very good” or higher. Four of those participants also reported having lived in a Spanish-speaking country for 8 or more years. Those participants were labeled as proficient Spanish speakers for the analysis.

All trials with response times of less than 200 ms were excluded from the analysis, as were trials with response times longer than 5466 ms, which was three standard deviations above the group's mean response time. A total of 83 trials were rejected based on these criteria, out of 9760 total trials. A total of 21 participants had at least one trial rejected, with the number of trials rejected ranging from 1-12.

In the response time analysis, there were 26 participants in the Spanish Proficient group and 33 participants in the Not Proficient group. The mean response times for each group for the match and mismatch priming conditions are shown in Figure 8. Only trials where the participants responded correctly were included in this analysis.

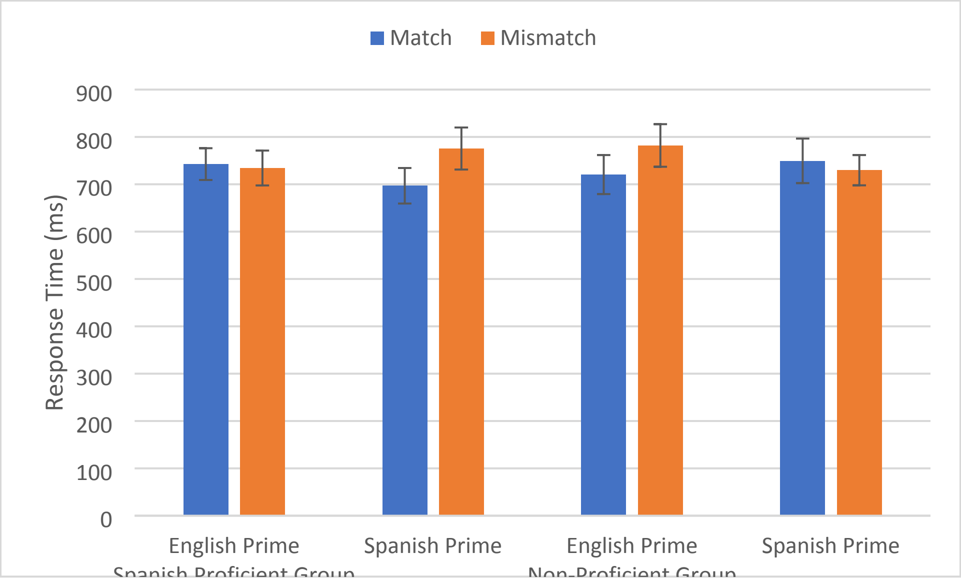

**Figure 8. The mean response times for match and mismatch primes in each language condition for each group of participants.**

A mixed 2x2x2 ANOVA with proficiency group, prime language, and priming condition as the factors showed that there was a significant three-way interaction between the three factors ($F(1,171) = 22.06$, $p < 0.001$). Paired t-tests comparing the match and mismatch priming conditions for each language within each group showed that the Spanish Proficient group had a significant priming effect for the Spanish primes ($t(25) = 5.73$, $p < 0.001$) but not for the English primes ($t(25) = 0.51$). Conversely the Non-Proficient group had a significant priming effect for the English primes ($t(32) = 3.38$, $p < 0.001$) but not for the Spanish primes ($t(32) = 0.91$).

## 4.3. Modeling Results

The response time data for all trials and the priming effects were modeled for this experiment using the same procedure that was used in the prior experiments. A random subset of the Non-Proficient participants were used in the modeling to make the sizes of the two groups of participants equal. When modeling the response time data, the model simply used each participant's response time for every trial. When modeling the priming effects, the model used the magnitude of the priming effect for each target picture in each language. This was done by subtracting the response times to each picture in the English Match condition from the response time in the English Mismatch condition, and then doing the same for the Spanish conditions. This produced two priming effect measures for each image, one for English and one for Spanish.

When using the response time data, the mean accuracy of the model was 68% correct (SD = 19%). The model performed about equally well at correctly categorizing the participants in the Spanish Proficient group (68% correct) and the Non-Proficient group (72% correct).

When using the priming effects rather than the response times for each trial, the model's accuracy dropped to an average of 48% correct (SD = 18%). There was a drop in accuracy for the Non-Proficient group, with the model only categorizing 61% of those participants correctly. There was an even larger drop in performance for the Spanish Proficient group, with the model only categorizing 35% of those participants correctly.

### 4.4. Discussion

The picture priming task successfully induced semantic priming in our participants, unlike the word length judgement tasks used in the prior experiments. The model performed reasonably well when making predictions based on the response times for all trials. However, the model's performance was somewhat lower in this task than it was in the first word length judgment task. It seems that the presence of a semantic priming effect did not improve the model's performance.

It is interesting to note that across all of the experiments, the model performed better when using response times for every trial rather than priming effects across conditions. In the prior experiments, where no robust priming effects were observed at the group level, we might expect the model to perform poorly when given only the priming effects. However, in the picture priming task, we observed significant priming effects at the group level and the model still performed poorly when using the priming data. It is possible that this superior performance for response times is due to the model having more data points to work with when every trial is considered separately. However, it is notable that this pattern was so consistent across experiments, and that it persisted even when there was a significant priming effect. This suggests that the model is picking up on something other than semantic priming effects in the patterns of response times that is predictive of language proficiency. Further research is needed to determine what factors other than semantic priming impact the model's performance.

# 5. MULTILINGUAL STROOP TASKS

## 5.1. Background

Given that semantic priming did not seem to improve the model's ability to predict which participants were proficient in different languages, we sought to test another well-established finding: semantic interference. To do this, we developed a series of multilingual Stroop tasks. Since its development in 1935, The Stroop task has been extensively used in a variety of domains to investigate selective attention, processing speed, and executive function (Stroop, 1935). In the classic color/word task, participants name the printed color of the word in either the congruent ink color (ex: GREEN printed word in green ink) or incongruent ink color (ex: GREEN printed word in red ink). In the majority of studies, incongruent trials elicit significantly longer reaction times, theorized to be due to cognitive interference of the competing information. This phenomenon is called the Stroop effect (for review see MacLeod, 1991). This is another task that participants can perform without being proficient in all of the languages in which the stimuli are presented. In the field of bilingualism research, the Stroop task has been used to study cognitive control, interference of words between and within languages, automaticity of access across different languages, and the possibility of a "bilingual advantage" in executive function (for review see van den Noort et. al., 2019).

### *5.1.1. Bilingualism and the Stroop effect*

In the bilingual Stroop task, color words can be presented in both the participants first language (L1) and second language (L2). This adds another dimension to the Stroop effect with the addition of between-language interference and within-language interference effects. Previous research has shown both interference types are affected differently depending on different factors with a pattern emerging of larger within-language Stroop effects (McLeod, 1991). Stroop interference is larger in response to words presented in the participants first language than other language words. This has been named the "within-language Stroop superiority effect" (Goldfarb & Tzelgov, 2007). These effects are modulated by factors such as age of acquisition (AoA) of L2 (Hernandez & Li, 2007, Sabourin, Vinerte, & Mayo, 2015), language proficiency of L2 (Zied et.al., 2004, Mägiste, 1984), as well as orthographic similarity between L1 and L2 (Chen & Ho,1986; Lee & Chan, 2000; Sumiya & Healy, 2004).

Some studies have shown the earlier a language is learned, the less interference in incongruent trials and the less difference in Stroop performance between L1 and L2 (Hernandez & Li, 2007, Tau et. al., 2011). Yow and Li (2015) found a positive relationship between AoA and interference costs. The earlier a person acquires L2, the better they performed on the Stroop task. Within early language learners, simultaneous learners and early sequential learners (1 – 6 years old) did not differ in Stroop performance when presented with one language, however, when their two languages were mixed, early learners exhibited facilitation of L1 and a larger interference effect for L2. This was not found for simultaneous learners (Sabourin & Vinerte, 2014).

More proficient bilinguals appear to better at inhibiting interference during the Stroop task than those that are less proficient (Zied, et. al 2004, Magiste, 1984, 1985, Chen and Ho, 1986, Okuniewska, 2007, Sutton, 2007, Tse and Altarriba, 2012, Singh & Mishra, 2012, Woumans et. al., 2015). Unbalanced bilinguals, those who are dominant in L1 and weaker in L2, exhibit larger Stroop effects in their first language than their second language, and experience greater interference from distractor was written in their dominate language. Balanced bilinguals (individuals who are equally proficient in both languages), on the other hand, do not show these differing effects for their different languages and have comparable interference for L1 and L2 (Mägiste, 1984, 1985; Chen and Ho, 1986). As L2 proficiency increases, overall reaction time on L2 Stroop trials decreases. The amount to which reaction times decreased in incongruent trials compared to congruent trials able differed with L2 proficiency (Tse & Altarriba, 2012).

Orthographic similarities/dissimilarities between languages may also impact bilingual Stroop performance. Researchers have found high between-language interference in highly proficient bilinguals when languages are similar (e.g., German-Swedish or German-English) while small between-language interference was found when languages are different orthographically (e.g., English-Greek or English-Chinese; Brauer, 1998, Magiste, 1992). This effect, or lack thereof, has also been found for Japan-English bilinguals, with Japanese also having smaller within-language interference when compared to English (Fang et. al., 1981). Conversely, Sumiya & Healy (2004) did find significant between-language interference despite the orthographic dissimilarity between Japanese and English, however, the between-language Stroop effect was larger with phonologically similar terms.

### 5.1.2. The Bilingual Advantage? Bilingual vs monolingual performance on the Stroop task

A large area of the research in bilingualism investigates the "bilingual advantage" in cognitive control and overall executive function (Bialystok et. al., 2003). Previous research has found that bilinguals outperform monolinguals on a variety of cognitive control tasking including the Stroop task (Bialystok,1999; Bialystok, Craik, & Luk, 2008; Costa, Hernandez & Sebastian-Galles, 2008). In the Stroop task, Bialystok and colleagues (2008) found that across age groups, bilinguals had greater interference suppression than monolinguals during the task. This effect was largest for older adults. This ability is postulated to arise from bilingual's ability to manage multiple languages a once.

There is also limited electrophysiological evidence to support differences, and perhaps advantages, in bilingualism when compared to monolingualism. A later N400 peak latency, possibly specific to L2, has been found for bilinguals during the Stroop task (Badzakova-Trajkov et. al., 2009; Coderre & Van Heuven, 2014a). N400 and LPC amplitude differences have also been found for congruent and incongruent trials in bilinguals but not in monolinguals (Heidlmayr et. al., 2015). Bilinguals also exhibit smaller N2 and Ninc during the Stroop task than monolinguals, an indication of suppression of interfering information (Coderre & van Heuven, 2013, Kousaie & Phillips, 2012a). Coderre and Van Heuven (2014b) also found a more negative amplitudes for L1 and no difference for L2 when compared to bilinguals. Other studies have not found this difference between L1 and L2 neural indices (Badzakova-Trajkov et. al., 2009; Heidlmayr, Hemforth, Moutier, & Isel, 2015).

More recently, the bilingual advantage has been called into question. More and more research in the field had found mixed or null results in support of the bilingual advantage (Paap & Greenberg, 2013, Kousaie & Phillips, 2012b, Coderra, Heuven & Conklin, 2013). Reviews in the area have revealed publication biases and small sample sizes to be possible underlying variables to attribute to this effect (Paap, Johnson, & Sawi, 2015, Bruin, Treccani, & Sala, 2015). Some studies have even found bilinguals perform worse on the Stroop task. Okada, He, and Gonzales (2019) found that young adult bilinguals had significantly slower reaction times during the taboo Stroop task than monolinguals. Other research has found that the variables that effect bilingual performance on the Stroop described earlier (AoA, proficiency, orthographic similarities between language) also play a role in comparing bilinguals to monolinguals (Yow and Li, 2015; Coderre & van Heuven, 2014).

While there is some evidence of a bilingual advantage, it could be reliant on these individual differences and be more task specific rather than the overall general advantages in executive functioning that have been theorized (van den Noort et. al., 2019).

Due to the long history of using the Stroop task to study bilingualism, we hypothesized that it might be effective for making predictions about language proficiency. In our prior experiments, our word length judgment tasks and picture priming task showed that our model performed better when using response time data for every trial rather than priming data calculated for the same stimuli in different conditions. With the Stroop tasks, our goal was to determine whether this pattern held in a task designed to induce interference rather than priming. We hypothesized that participants would be slower to respond to incongruent trials in languages that they understand, but would be unaffected by incongruent trials in languages that they do not understand. Furthermore, we predicted that both the response times for every trial and the Stroop interference effects could be modeled to make predictions about which participants understood which languages.

We tested two versions of the Stroop task. The first used only three colors: red, green and blue. The color words were presented in English, Spanish, French and German. However, some people may be familiar with the basic color words in one or more of those languages, even if they are not proficient in the language in general. Because of this, we also developed a nine color version of the Stroop task to determine whether a broader set of color names would lead to more predictive results.

## 5.2. Three Color Stroop Experiment

### *5.2.1. Methods*

#### 5.2.1.1. Participants

A total of 67 participants completed the experiment on Amazon Mechanical Turk. All participants were required to have an approval rating greater than 90% for prior tasks completed on AMT. Thirty-two of the participants had AMT's Spanish fluency qualification, four had the French fluency qualification, and one had the German fluency qualification. Another 30 participants completed a version of the task that did not require any specific language fluency qualifications (although some of the participants in this group may have had those qualifications).

#### 5.2.1.2. Materials

The materials for this task consisted of three color words that were presented in English (red, green, blue), Spanish (rojo, verde, azul), French (rouge, vert, bleu), and German (rot, grün, blau). There was also a control condition, predicted to have no interference effect, in which participants saw only the letters "xxxx." The control condition was presented nine times in the experimental list, three times in each of the three possible font colors (red, green and blue). Each of the other words appeared six times in the list, three times in the congruent condition and three times in the incongruent condition. In the congruent condition, the word was shown in a font color that matched its meaning (i.e., **red**). In the incongruent condition, the word was shown in one of the non-matching font colors (i.e., **red** or **red**). The list contained a total of 81 items.

#### 5.2.1.3. Procedure

After acknowledging the consent form, the participants completed an abbreviated version of the language history survey that was used in the other tasks. The survey asked them to list the languages they know in order of dominance, the languages they know in order of acquisition, the percentage of time they are currently exposed to English, Spanish, French and German (on average), and to rate their level of proficiency in reading English, Spanish, French and German on the same scale that was used in the prior experiments.

Next they read the instructions screen, which told them to press the button corresponding to the color of the font, not the meaning of the word. They were instructed to answer as quickly as possible. When they were ready, the participants clicked a button on the screen to begin the task. Unbeknownst to the participants, the first six trials were practice trials that were not included in the analysis. They gave participants a chance to become familiar with the task and the layout of the three response buttons. On each trial, the color word was shown in 20-point bold Arial font, centered above three response buttons which were labeled "red," "green" and "blue." Figure 9 shows an example of how the trials looked to the participants.

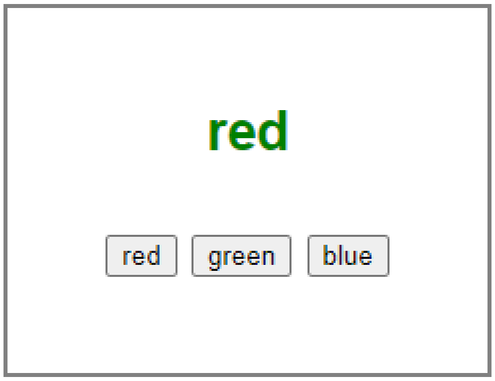

**Figure 9. An example of a trial in the three color Stroop task.**

The word stayed on the screen until the participant clicked on one of the buttons. Then it was replaced by the next word in the list. The six practice trials were followed immediately by the 81 real trials, which were presented in a different random order for each participant.

### *5.2.2. Results*

One participant was removed from the analysis because they reported low proficiency in English. Among the remaining participants, one person rated their proficiency in reading English as being "good" and all of the other participants rated their proficiency as "very good" or better. There were 26 participants who rated their Spanish proficiency as "very good" or better. This included all of the participants who had AMT's Spanish fluency qualification, plus four participants who completed the version of the task that did not require any specific language fluency qualifications. There were five participants who rated their proficiency in French as being "very good" or higher, including all of the participants who had AMT's French fluency qualification and one participant who had the Spanish fluency qualification. Only one participant, the participant who had AMT's German fluency qualification, rated themselves as being proficient in German.

Prior to analysis, all trials with response times of less than 200 ms were removed, as were trials with response times longer than 16601 ms, which was three standard deviations above the group mean response time. A total of 13 trials from five participants were rejected.

The response time results for all trials on which participants responded correctly are shown in Figure 10. For the Spanish Proficient group, paired t-tests showed that there was a Stroop effect, with significantly shorter response times for congruous than for incongruous trials, for the English word condition ($t(25) = 3.53$, $p < 0.001$) but not for the Spanish ($t(25) = 1.29$, $p = 0.10$), French ($t(25) = 1.71$, $p = 0.05$), or German ($t(25) = 0.81$) conditions. For the Non-Proficient group, there was a significant Stroop effect for the English ($t(29) = 3.34$, $p < 0.01$), Spanish ($t(29) = 2.31$, $p < 0.02$) and German ($t(29) = 2.10$, $p < 0.03$) conditions, but not for the French condition ($t(29) = 0.19$).

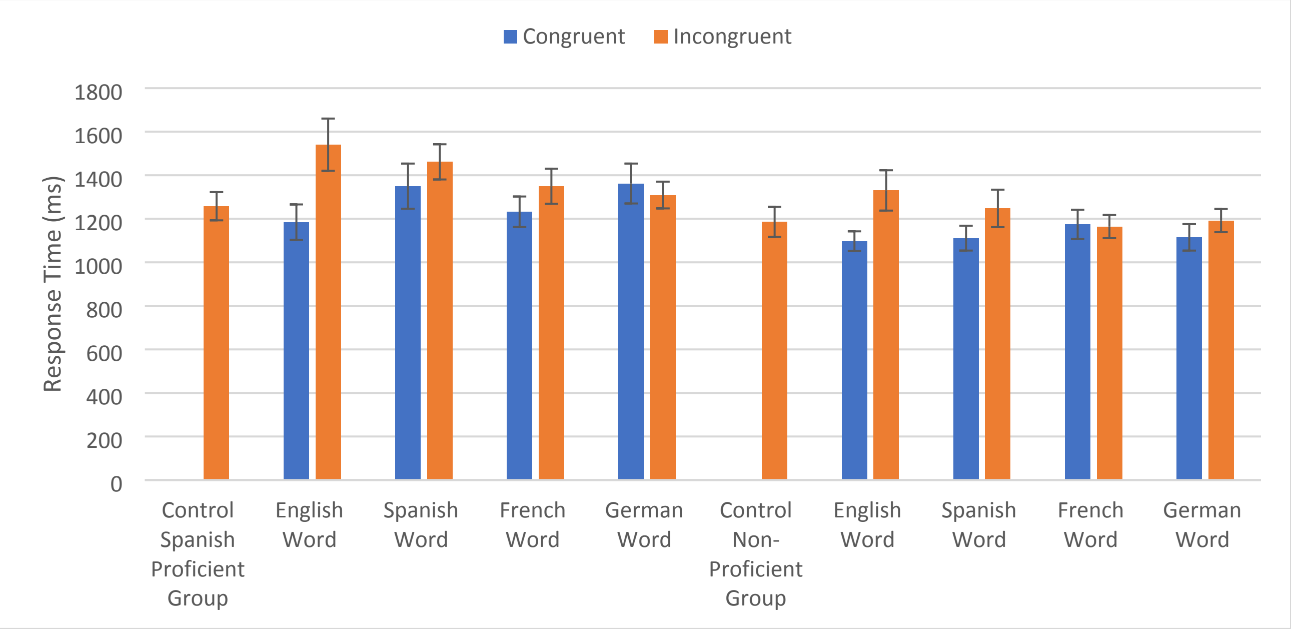

**Figure 10. Average response times for each condition for the Spanish Proficient and Non-Proficient Groups.**

When we looked only at the participants who were fluent in French, we observed a large Stroop effect for the French stimuli for those participants (mean Stoop effect size = 217.5 ms for these five participants). Similarly, our lone German participant showed a Stroop effect for the German stimuli (404.8 ms). Thus, it was surprising that the proficient Spanish speakers did not show a significant Stroop effect in Spanish. Individual participants in this group showed a large Stroop effect, but a handful of participants had a very small effect or a negative effect.

It was also surprising that the group of participants who were not proficient in Spanish showed significant Stroop effects for Spanish and German. The size of the Stroop effect was smaller on average for those languages, but the difference between congruent and incongruent conditions was still significant.

### 5.2.3. Modeling Results

The participants' response times for each trial were used to predict whether or not they were proficient in Spanish, using the method described in Experiment 1. Since there were so few participants who were proficient in French or German, we did not attempt to model those languages. Participants who had the French or German fluency criterion, or who reported fluency in French or German, were simply labeled for the modeling based on their responses to the questions about English and Spanish proficiency. This gave us a group of 31 participants who were proficient in English. To match that sample size, a group of 31 people who were proficient in both English and Spanish were randomly selected from among the 35 such participants in our dataset.

The model's performance was close to chance in its predictions about which participants were proficient in both English and Spanish. When using the response times for every trial, the model had a mean accuracy of 52% (SD = 10%). When using averages across conditions (congruent versus incongruent trials), the model performed somewhat better, with a mean accuracy of 57% correct (SD = 13%). In both versions of the model, it correctly categorized 57-58% of the participants who were not proficient in Spanish. The version that used the response times for each trial was quite poor at identifying the participants who were proficient in both languages, correctly identifying them only 49% of the time. The version that used averages across conditions performed better for the bilingual participants, reaching 60% accuracy. However, the model's performance was still quite modest in both cases.

### 5.2.4. Discussion

The results from the three color Stroop task were somewhat surprising. The Spanish Proficient group did not show a significant Stroop effect for Spanish stimuli, while the participants in the Non-Proficient group did. Unsurprisingly, given these results, the model did not perform well when attempting to predict whether or not individual participants were proficient in Spanish.

It is possible that the unexpected results from this task were caused by the online implementation. The participants had to move the mouse to click on a button, which is very different from the traditional implementation of the Stroop task where people say the color of the font out loud. It is also possible that many of our participants were familiar with the color words in Spanish, French, and/or German, even if they were not proficient in those languages. The words for "red," "green" and "blue" are quite similar across these languages, with the exception of "azul" in Spanish. The similarities between the languages or the participants' familiarity with these specific color words could have had an impact on the results as well.

Due to the COVID-19 pandemic, we could not implement a version of this task in which participants spoke the color names out loud, so we were unable to address the first potential issue. However, in an attempt to address the second issue, we developed a nine color version of the Stroop task. We suspected that fewer people would be familiar with less common color words in Spanish, French, and German. In addition, many of these color words were not visually or auditorily similar to their English equivalents. We hypothesized that using a larger set of color words might produce better data for modeling purposes.

## 5.3. Nine Color Stroop Experiment

### *5.3.1. Methods*

#### 5.3.1.1. Participants

A total of 45 participants completed the experiment on AMT. All participants were required to have an approval rating greater than 90% for prior tasks completed on AMT. Twelve of the participants had AMT's Spanish fluency qualification, two had the French fluency qualification, and one had the German fluency qualification. Another 30 participants completed a version of the task that did not require any specific language fluency qualifications (although some of the participants in this group may have had those qualifications).

#### 5.3.1.2. Materials

The materials for this task consisted of nine color words that were presented in English (red, yellow, green, blue, purple, pink, brown, white, gray), Spanish (rojo, amarillo, verde, azul, morado, rosado, marrón, blanco, gris), and German (rot, gelb, grün, blau, lila, rosa, braun, weiß, grau). There was also a control condition, predicted to have no interference effect, in which participants saw only the letters "xxxx."

In this version of the task, French was not used as one of the languages because we had few responses from participants with the French fluency qualification in other experiments. To keep the task a reasonable length, we chose to eliminate one of the languages. We made the assumption that few participants would be fluent in German, allowing it to serve as a second control condition with real words as stimuli.

The "xxxx" items were presented 18 times in the experimental list, twice in each of the nine possible font colors. Each of the other words appeared four times in the list, twice in the congruent condition and twice in the incongruent condition. In the congruent condition, the word was shown in a font color that matched its meaning. In the incongruent condition, the word was shown in one of the non-matching font colors. The list contained a total of 126 items.

#### 5.3.1.3. Procedure

The procedure used for this experiment was identical to the procedure for the three color Stroop task, except that the question about how many years and months the participants had spent living in predominantly English, Spanish, French, and German speaking countries was added back in to the language history questionnaire.

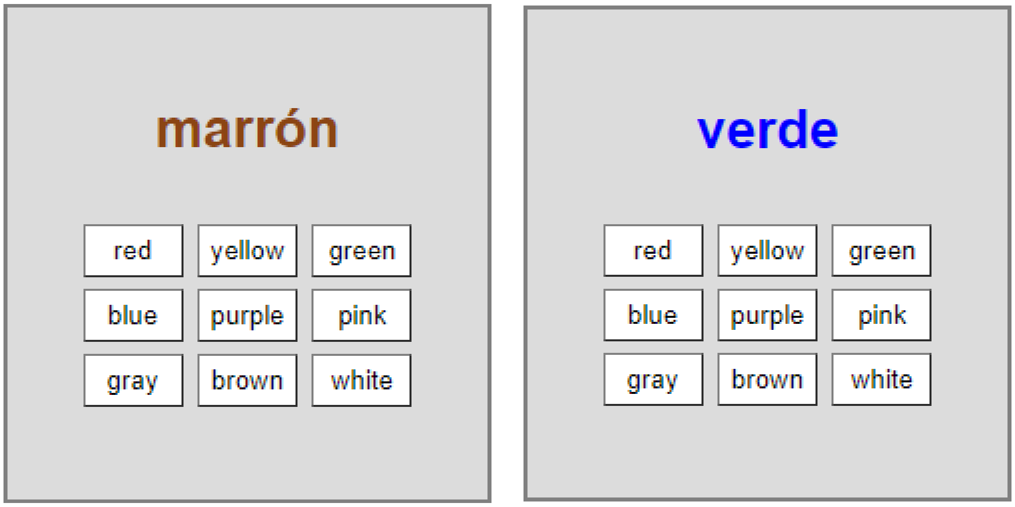

**Figure 11. Examples of a congruent trial (left) and an incongruent trial (right), both using Spanish words, in the nine color Stroop task.**

### *5.3.2. Results*

Two participants were removed from the data analysis because they did not provide valid responses to the language survey. In the group that had AMT's Spanish fluency qualification, all of the participants reported a high level of proficiency in both Spanish and English. The participants in this group had lived an average of 29 years in predominately Spanish-speaking countries (range = 1-49 years) and an average of 12 years in predominately English-speaking countries (range = 0-40 years).

The participants who had AMT's French or German fluency qualification all reported high proficiency in English, but not in Spanish, so they were grouped with the non-proficient Spanish speakers for the analysis. Among the non-proficient Spanish speakers, only one person reported having a "good" level of proficiency in Spanish. That same person reported having lived in a Spanish-speaking country for two years. None of the other participants in this group reported spending any time living in a Spanish-speaking country. They had spent an average of 34.6 years living in English-speaking countries (range = 0-59).

Prior to analysis, all trials with response times of less than 200 ms were removed, as were trials with response times longer than 9435 ms, which was three standard deviations above the group mean response time. A total of 16 trials from five participants were rejected.

The response time results for all trials on which participants responded correctly are shown in Figure 12. For the Spanish Proficient group, paired t-tests showed that there was a Stroop effect, with significantly shorter response times for congruous than for incongruous trials, for the English condition ($t(11) = 2.88$, $p < 0.01$) but not for the Spanish ($t(11) = 0.95$) or German ($t(11) = 0.16$) conditions. For the Non-Proficient group, there was a significant Stroop effect for the English ($t(30) = 5.02$, $p < 0.001$) and German ($t(30) = 4.05$, $p < 0.001$) conditions, but not for the Spanish condition ($t(30) = 1.46$, $p = 0.08$).

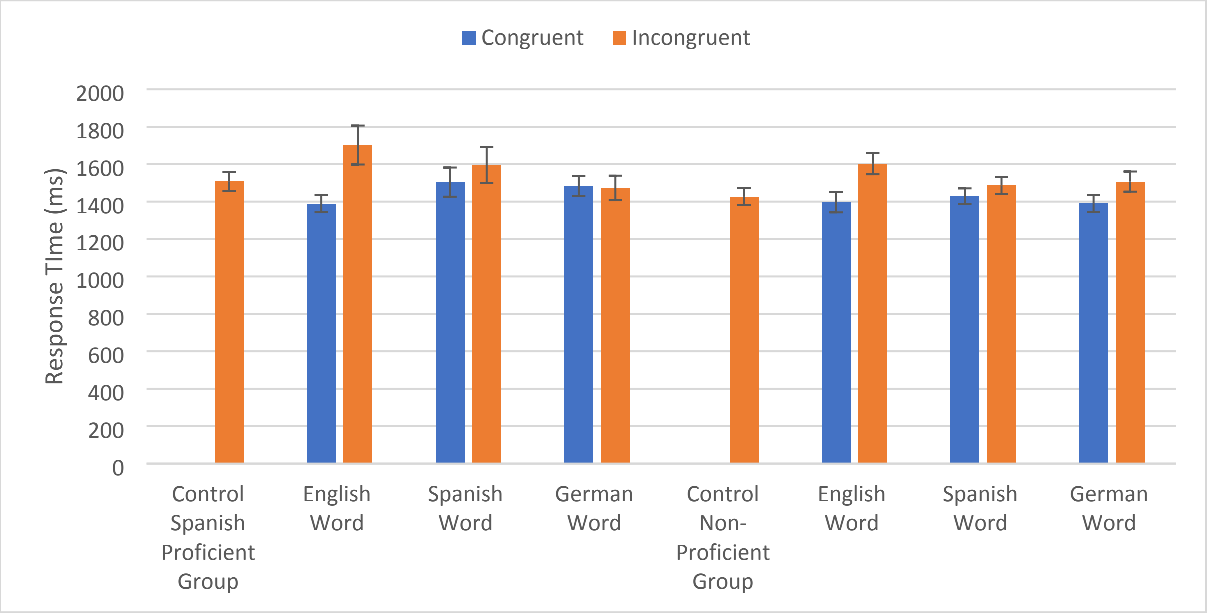

**Figure 12. Average response times in each condition for the Spanish Proficient and Non-Proficient Groups.**

The Stroop effect was not significant for the proficient Spanish speakers in this study, mirroring the results of our three color Stroop task. Similarly, the non-proficient speakers showed a significant Stoop effects in English and German, just like they did in the three color Stroop task. However, in this case they did not show a significant Stroop effect for Spanish. While there was one proficient German speaker in that group, the German Stroop effect was still significant even if that participant's data were removed, so that person alone was not driving the effect. These results indicate that the nine color Stroop task was unlikely to be effective in distinguishing people with proficiency in different languages.

### 5.3.3. Modeling Results

The data was modeled using the same process that was used for the three color Stroop task. As expected, the model performed very poorly. When using the response times for every trial, the model had a mean accuracy of 35% (SD = 19%). When using averages across conditions (congruent versus incongruent trials), the mean accuracy was 31% correct (SD = 15%).

### 5.3.4. Discussion

While we had anticipated that using a larger number of color words would allow us to better discriminate between monolingual and bilingual participants, that did not turn out to be the case. The model's performance was substantially worse for the nine color Stroop task than it was for the three color Stroop task. It is possible that in the nine color task, the additional time needed to locate the correct response button washed out some of the effects. This suggests that our online implementation of this task was not successful.

In both the three color and nine color Stroop tasks, our Spanish Proficient groups showed Stroop effects for English but not for Spanish. It is possible that this pattern reflects the bilingual advantage that has been observed in the prior literature, where bilingual participants show reduced interference effects in their L1. However, even though this finding was consistent across the two experiments, this pattern was not useful for the purposes of making predictions about individual participants' language proficiency. Many of the monolingual English speakers also showed a Stroop effect in English but not Spanish. When looking at the participants as individuals rather than looking at group level data, seeing a Stroop effect for English but not for Spanish could indicate one of two things. First, it could indicate that the participant did not understand Spanish and thus did not suffer from any interference from the Spanish words. Alternatively, it could indicate that the participant was a proficient bilingual who showed reduced interference in Spanish due to the bilingual advantage. With two possible explanations for the same pattern of effects, it is not surprising that the ML approach failed to make good predictions. In the online implementation of this task, any subtle differences that might have allowed us to distinguish between these two possibilities were lost.

## 5.4. Stroop Experiment with Non-Roman Characters

In addition to testing the Stroop effects in English, Spanish, French and German, we also collected data in a task that incorporated languages that use non-Roman characters. In this version of the task, the languages used were English, Hindi, Mandarin, and Irish (Gaelic). Hindi and Mandarin were chosen because of their writing systems and because AMT has fluency qualifications for those languages. Irish was chosen because the color words look similar to English words, but we assumed that there would be very few proficient speakers of Irish in the AMT worker community. This gave us a condition that had Roman characters but words that would be unfamiliar to most of our participants.

This task mirrored the three color Stroop task in using red, green, and blue as the font colors. The color words were shown in English (red, green, blue), Irish (dearg, glas, gorm), Mandarin (红色, 绿色，蓝色）and Hindi (लाल, हरा, नीला). The structure of the task was identical to the structure of the three color Stroop task.

Twenty-eight people participated in the version of the task that did not require any language proficiency qualifications. Only one participant with AMT's Mandarin fluency qualification completed the experiment, but three of the participants in the group that were not required to have any language fluency qualifications reported that they had a "good" level of proficiency with reading Mandarin (a 7 on the 0-10 scale). However, those same three participants also reported moderate to high levels of proficiency in Irish and Hindi, which calls their survey results into question.

There were no participants in the version of the task that required AMT's Hindi fluency qualification, and only one participant in the unrestricted version of the task who reported that they had "perfect" proficiency in reading Hindi. However, in the free response portion of the survey that came prior to the language proficiency ratings, this participant reported that English was the only language they knew. So once again, this person's survey results were suspect.

Unsurprisingly, given the lack of participants with proficiency in Hindi, Mandarin, or Irish, we did not observe Stroop effects in those languages (all *t*s < 1.46, all *p*s > 0.08). There was a Stroop effect for the English words, with participants taking an average of 263 ms longer to respond to incongruent words than to congruent words ($t(28) = 5.09$, $p < 0.001$). These results are shown in Figure 13.

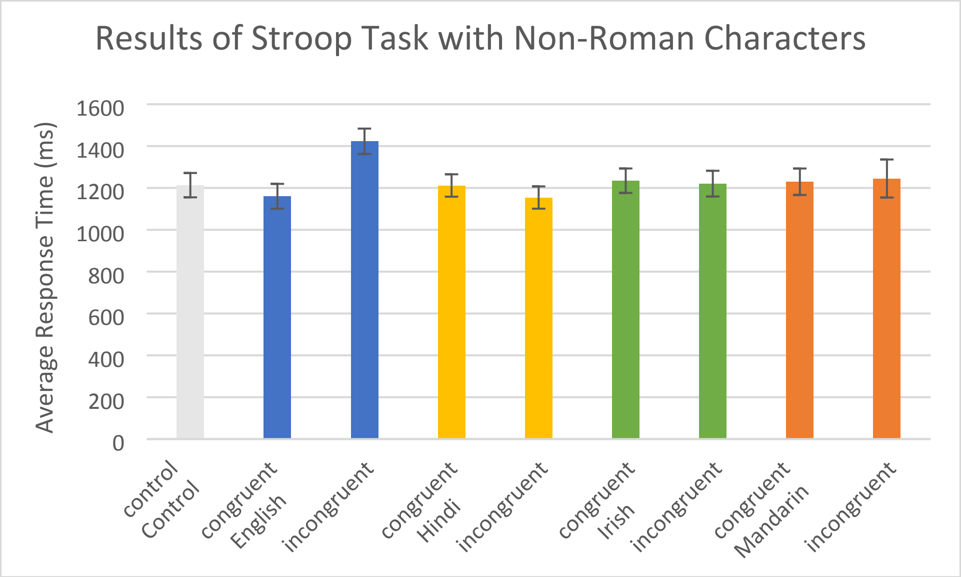

**Figure 13. Average response times to congruent and incongruent stimuli in each language.**

Looking at the only participant who had AMT's Mandarin fluency qualification, we found that that person showed a Stroop effect in both English (112 ms) and Mandarin (386 ms). However, they also showed a Stroop effect in Irish, a language with which they were not familiar (163 ms).

Due to the lack of participants who were fluent in Hindi or Mandarin, we did not model the results of this study. Overall, this task provided a case study in some of the pitfalls of collecting data online. In some cases, participants submitted invalid responses to the surveys or pressed buttons randomly during the tasks. On this and all other tasks, we had to be careful to remove participants whose responses seemed random or whose survey results were implausible or internally inconsistent.

# 6. ELECTROENCEPHALOGRAPHY (EEG) STUDY

## 6.1. Background

EEG has been used to study the electrical activity of the human brain for nearly 100 years. As sensor technology and computing power have advanced, EEG recording techniques have provided increasingly fine-grained insights into how the brain processes information. In the mid 1960s, researchers developed techniques for time-locking EEG signals to the onset of stimuli, allowing them to isolate neural signals that were related to the brain's processing of those stimuli (Sutton, Braren & Zubin, 1965; Walter et al., 1964). The time-locked EEG signals are called event-related potentials (ERPs). Since the discovery of ERPs, thousands of studies have focused on characterizing these signals and their relationships to specific neural processes. ERPs are named according to their polarity and their timing (in milliseconds) relative to the onset of the stimulus.

In 1980, researchers discovered the N400, an ERP that was elicited by verbal stimuli (Kutas & Hillyard, 1980). Subsequent decades of research have determined that the N400 reflects semantic processing (i.e., the processing that takes place when the brain accesses the meaning of a word, or attempts to access the semantics of any potentially meaningful stimuli). Further research identified another ERP, the P600, which reflects the brain's processing of the structure of language. If a person encounters a grammatical error, their brain produces a P600 as it tries to reanalyze the structure of the sentence (Osterhout et al., 2008).

These two ERPs related to language processing are extremely well-characterized and they are highly consistent across individuals. The brain processes language automatically, so the N400 and P600 can provide detailed information about a person's knowledge without any overt response required from the person. For example, the N400 can distinguish the words a person knows from words he or she does not know (Kutas & Federmeier, 2011). The N400 and P600 have been shown to track proficiency when a person is learning a second language (McLaughlin et al., 2004), to indicate the age at which a person acquired a language (Weber & Lavric, 2008), and even to reflect cultural biases (a P600 is elicited by pronouns that violate cultural gender norms; Oakhill et al., 2005). These components can also be used to assess a person's retention of newly-memorized verbal information (Haass & Matzen, 2011), to understand the disruption to comprehension caused by misspelled words (Stites, Federmeier, & Christianson, 2016), or to predict a developing reader's future reading abilities (Stites & Laszlo, 2017).

The academic research on ERPs has focused on studying groups of people with specific characteristics and then extrapolating to generalized principles about neural processing. To make use of this body of literature in more applied settings, we must do the opposite: apply general principles of neural processing to learn about individual people. Thus, although there is a long and robust history of research in this area, there are key questions that have not been addressed by the academic community. In a typical ERP study, researchers recruit participants who meet specific criteria, then examine how group differences relate to differences in ERP amplitude, timing, and/or scalp distribution. For example, in the existing studies that link the N400 and P600 to progress in learning a new language, researchers recruited students who were taking entry-level language classes at their university and tracked them over the course of a semester. All of the other studies that have demonstrated that the N400 or P600 were sensitive to differences such as language proficiency knew how the participants differed ahead of time and lumped them into group-level analyses rather than assessing each individual separately. In order to apply this research in a new way, we need to address the following question: Can ERPs identify which languages an individual understands when the researcher does not know ahead of time?

To address this question, we designed an ERP study that built upon on existing, well-established ERP experiment protocols from the academic literature. Participants were shown a list of related and unrelated word pairs in English, Spanish, French and German. The related word pairs either had the same word repeated twice, or a word in one language paired with its translation in another language. In the unrelated pairs, two semantically unrelated words were shown, either in the same language or in two different languages. For languages that a person understands, we would expect to see an N400 priming effect where the N400 amplitude is reduced when the participant reads the second word in a related pair. The N400 priming effect is extremely robust and has been observed in cross-language experiments (for a review, see Moreno, Rodriguez-Fornells, & Laine, 2008). For languages that the participant does not understand, their brain would not access the meaning of the second word in the pair and therefore they would not exhibit an N400 priming effect.

Although this basic semantic priming paradigm has been widely used in the ERP literature, our experiment differed from the prior research in three important ways. First, we presented the participants with stimuli from multiple languages (prior studies have used two at most and have explicitly selected participants who were known to be proficient with both languages). Second, we intermixed the languages instead of grouping targets from the same language into block. Third, we made the experiment blind so that the experimenter analyzing the data did not know which language(s) each participant understands. The data analyst was tasked with predicting which languages each participant understands based on the ERP data alone, with no other information about the participant provided. In addition, we attempted to make predictions about language proficiency using a machine learning model based on the participants' ERP data. Our goal was to be able to predict which language(s) each participant understands at a level greater than chance.

## 6.2. Methods

### 6.2.1. Participants

Forty employees of Sandia National Laboratories participated in this experiment. They were compensated for their time at their usual hourly rate. Of the participants, 21 were female and 19 were male. Their average age was 39.6 years (range = 23-68). There was a range of levels of education among the participants. One participant reported "less than high school" as their highest degree obtained. Four participants reported "some college" or a bachelor's degree. One participant reported "some graduate school." Nineteen participants reported having a Master's degree and 14 of the participants reported having a Ph.D.

### 6.2.2. Materials

The materials consisted of 30 English nouns and their translations in Spanish, French, and German. The translations were reviewed by people who were native or highly proficient speakers of Spanish, French and German to ensure that the translations were accurate. The words were selected so that there were no special characters (accents, umlauts, etc.), and no cognates (e.g., family, familia) or false cognates (e.g., soap, sopa) across the four languages. Since the participants in the experiment live in Albuquerque, New Mexico, where many place and street names are Spanish words, we took care to select Spanish words for our experiment that monolingual English speakers would be unlikely to encounter during their daily lives in Albuquerque. The stimuli are shown in Table 3.

**Table 3. The stimuli for the EEG experiment.**

| English Words | Spanish Words | French Words | German Words |
|---|---|---|---|
| BOOK | LIBRO | LIVRE | BUCH |
| CHAIR | SILLA | CHAISE | SESSEL |
| CITY | CIUDAD | VILLE | STADT |
| CLOUD | NUBE | NUAGE | WOLKE |
| DAUGHTER | HIJA | FILLE | TOCHTER |
| FACE | CARA | VISAGE | GESICHT |
| FLOUR | HARINA | FARINE | MEHL |
| FOOD | COMIDA | NOURRITURE | ESSEN |
| GARLIC | AJO | AIL | KNOBLAUCH |
| GROUND | SUELO | SOL | BODEN |
| HAPPINESS | FELICIDAD | BONHEUR | FREUDE |
| KNIFE | CUCHILLO | COUTEAU | MESSER |
| LAUGHTER | RISA | RIRE | LACHEN |
| LAWYER | ABOGADO | AVOCAT | ANWALT |
| LEAF | HOJA | FEUILLE | BLATT |
| MOUTH | BOCA | BOUCHE | MUND |
| NECK | CUELLO | COU | GENICK |
| NEIGHBOR | VECINO | VOISIN | NACHBAR |
| NEPHEW | SOBRINO | NEVEU | NEFFE |
| PILLOW | ALMOHADA | OREILLER | KISSEN |
| PLACE | LUGAR | ENDROIT | ORT |
| POCKET | BOLSILLO | POCHE | TASCHE |
| RAIN | LLUVIA | PLUIE | REGEN |
| ROOM | SALA | CHAMBRE | ZIMMER |
| SHOE | ZAPATO | CHAUSSURE | SCHUH |
| SKIN | PIEL | PEAU | HAUT |
| WALL | PARED | MUR | MAUER |
| WATCH | RELOJ | MONTRE | UHR |
| WOOD | MADERA | BOIS | HOLZ |
| YESTERDAY | AYER | HIER | GESTERN |

The word lists were matching on length, frequency, and orthographic neighborhood size. This information was acquired from the CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012). The average length, frequency (per million), and orthographic neighborhood size for each language is shown in Table 4. We tried to minimize cross-language orthographic neighborhood size, which is less than four for all of the language pairings.

Table 4. The psycholinguistic properties of the stimuli in each language.

| Language | Avg Length | Avg Frequency | Avg English Orthographic Neighborhood Size | Avg Spanish Orthographic Neighborhood Size | Avg French Orthographic Neighborhood Size | Avg German Orthographic Neighborhood Size |
|---|---|---|---|---|---|---|
| **English** | 5.37 | 106.93 | 6.77 | 0.77 | 2.30 | 1.77 |
| **Spanish** | 5.47 | 99.04 | 2.27 | 5.43 | 1.43 | 1.47 |
| **French** | 5.57 | 101.39 | 3.87 | 1.27 | 6.80 | 2.67 |
| **German** | 5.40 | 94.39 | 2.10 | 0.40 | 0.90 | 5.40 |

### *6.2.3. Word Pairs*

Each English word was paired with itself and with its translation in the other three languages. Each of the words in Spanish, French and German was also paired with itself. For the sake of conciseness, we refer to these pairings collectively as the "translation pairs," even though a subset of the pairs are actually repetitions of the same word in the same language. Every word was also paired with unrelated pairs in the same language and in English. We refer to these pairings as the "unrelated pairs." The order of the languages in the pairings occurred in both possible orders, with English words appearing as the first word in the pair for half of the pairs and as the second word in the pair for the other half. Working through all of the possible pairings, this produced 20 conditions, as shown in Table 5.

Table 5. All possible pairings within and across languages.

| Translation Pairs | Unrelated Pairs |
|---|---|
| English-English Repetition | English-English Unrelated |
| English-Spanish Translation | English-Spanish Unrelated |
| English-French Translation | English-French Unrelated |
| English-German Translation | English-German Unrelated |
| Spanish-Spanish Repetition | Spanish-Spanish Unrelated |
| Spanish-English Translation | Spanish-English Unrelated |
| French-French Repetition | French-French Unrelated |
| French-English Translation | French-English Unrelated |
| German-German Repetition | German-German Unrelated |
| German-English Translation | German-English Unrelated |

All of the words were rotated through all of the possible pairings. When creating the unrelated pairs, we took care to avoid pairings that had a semantic relationship, such as "neck" and "mouth." In the end, there were 600 word pairs with every word appearing in every possible condition.The word pairs were divided into 10 counterbalanced blocks containing 60 pairs each. Each block contained three pairs corresponding to each of the pair types shown in Figure 3. The first word of the pair was in English for 24 of the pairs and in Spanish, French and German for 12 pairs each. The same was true for the second word in each pair. Within each block, each English word and/or its various possible translations appeared four times, twice in a translation pair (as the first and second word in the pair), once in an unrelated pair as the first word in the pair, and once in an unrelated pair as the second (target) word in the pair.

The pairs were placed in a pseudorandom order so that the two pairs containing each target word appeared in different halves of the block. This was counterbalanced so that any given word was the target word in a translation pair in the first half of five of the blocks and in the second half of the other five blocks, and vice versa for the same word serving as a target in unrelated pairs. Additionally, word pairs with the same target word never appeared in the same position in different blocks (i.e., if a target word was in the second pair in one block, it would not be in the second pair in any other blocks). Finally, the pseudorandom order was constrained so that there were never more than three translation or unrelated pairs in a row, and never more than two pairs from the same category (e.g., Spanish-English translation) in a row.

The counterbalancing was done so that the structure of the cross-language unrelated pairs could not be predicted from the related pairs in a given block. For example, each block had three English-English repetition pairs, and each of the target words in those pairs appeared with a different language when it was the target of an unrelated pair. One was preceded by an unrelated Spanish word, one by an unrelated French word, and one by an unrelated German word. There was no way for the participants to predict which English target word would be paired with which language. This also ensured that every Spanish, French and German target word was preceded equally often by a word in the same language or a word in English, while every English target word was preceded equally often by words in English, Spanish, French and German.

### *6.2.4. Response words*

The participants were tasked with responding via a button press whenever they saw an animal name in any language. They were instructed that the words “CAT” and “DOG,” as well as their translations in Spanish, French, and German, would be interspersed with the other words in the list. They were trained on the translations of both words in the other languages (GATO/CHAT/KATZE and PERRO/CHIEN/HUND) to ensure that they would be familiar with the to-be-responded-to words in every language. Participants were also told that other animal names might appear in any language and that they should press a button as soon as the recognized any animal name.

There were five additional animal names that appeared within the word lists: COW/VACA/VACHE/KUH, HORSE/CABALLO/CHEVAL/PFERD, SHEEP/OVEJA/MOUTON/SCHAF, DEER/CIERVO/CERF/HIRSCH, and RABBIT/CONEJO/LAPIN/HASE. These words were selected using the same criteria as the stimulus lists. These words had similar average lengths and frequencies across all four languages. These additional, untrained response words were intended to keep the participants alert and to provide a behavioral signal of which languages each participant understood. The trained and untrained response words were dispersed among the pairs in the word lists, with the gaps between the response words ranging from zero pairs (two response words presented back-to-back) to eight intervening pairs. On average, one response word appeared for every 3-4 word pairs. The all variants of the CAT and DOG appeared twice in each block, while two of the untrained response words appeared in each block.

**Table 6. The psycholinguistic properties of the animal names used for response trials.**

| Language | Avg Length | Avg Frequency | Avg English Orthographic Neighborhood Size | Avg Spanish Orthographic Neighborhood Size | Avg French Orthographic Neighborhood Size | Avg German Orthographic Neighborhood Size |
|---|---|---|---|---|---|---|
| **English** | 4.60 | 32.29 | 13.20 | 0.80 | 2.00 | 2.80 |
| **Spanish** | 5.60 | 23.39 | 0.20 | 4.00 | 0.00 | 0.00 |
| **French** | 5.20 | 33.01 | 1.00 | 0.80 | 3.20 | 1.40 |
| **German** | 4.60 | 19.49 | 4.00 | 1.60 | 2.00 | 6.60 |

### 6.2.5. Procedure

After completing the consent form, participants completed a demographics questionnaire, a handedness questionnaire, and the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al., 2007). To complete the main task, they were seated in a dimly lit sound booth with their eyes approximately 60 cm from a computer monitor. The experiment began with a training task in which the participants were trained on the words for cat and dog in Spanish, French and German. Then they completed a short practice list that mirrored the structure of the real task. The participants were instructed to press a button on a game controller any time they saw an animal name, whether it was one of the words they had just memorized or a different animal name. The practice list contained 28 nouns, including the words "cat" and "dog" in all four languages. There were also four other animal names, one in English and three that were cognates of English animal names (e.g., tigre, elefant). At the end of the practice list, the participants were told how many of the animal names they successfully identified. They were given the opportunity to repeat the practice list if they desired. Otherwise, they continued to the main experiment.

During the main experiment, each participant saw one of the experimental lists. The lists were divided into 10 blocks and the participants were given self-paced breaks between the blocks. The participants pressed a key on the computer keyboard to initiate each block, to ensure that they did not initiate the blocks accidentally. All participants responses during the blocks themselves were made using the game controller, which participants held in their laps. At the beginning of each block, the words "GET READY" appeared in red font on a black background for three seconds. Then the trials for that block began.

Each trial began with a fixation cross that was presented on the screen for 1000 ms. This was followed by a blank screen that was presented for 400 ms, and then the word, which appeared in the same location as the fixation cross and was presented for 1000 ms. Throughout all of the trials, the background of the screen was black. The fixation cross and words were presented in the center of the screen in silver 72-point Consolas font. At the end of each block, participants were told how many blocks they had completed and how many were remaining. Each block lasted for approximately five minutes.

The participants were instructed to sit still and to try not to blink when the words were on the screen. They pressed the response button on the game controller when they recognized an animal name in any languages. They did not respond to any of the other words.

### *6.2.6. EEG Recording and Data Analysis*

The ongoing encephalogram (EEG) was recorded from 32 active silver/silver-chloride electrodes using the ANT Neuro waveguard cap arranged in the 10/20 layout. Electrodes were referenced online to a ground electrode near Cz and were re-referenced offline to the average of the left and right mastoids. A bipolar eye channel was created by placing electrodes on the left infraorbital ridge and above the left eye, referenced to each other, to monitor for blinks. A second bipolar eye channel was created by placing electrodes on the outer canthus of each eye, referenced to each other, to monitor for horizontal eye movements. Impedances for scalp electrodes were kept below 25kΩ. The continuous EEG was recorded to hard disk at a sampling rate of 250Hz.

All data processing was completed using the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes for MATLAB. Epochs of EEG data were taken from 100 ms before stimulus onset to 900 ms post-stimulus. Those containing artifacts from signal drift, eye movements, eye blinks, muscle activity, or other types of noise were rejected off-line before averaging, using thresholds selected for each participant through visual inspection of the data. A two-step approach was used for artifact rejection. First, a simple voltage threshold filter was applied to all scalp channels, with individually-set thresholds for each participant. Second, a moving-window peak-to-peak filter was applied to one or both of the bipolar eye channels, with a window size of 175 ms and a window step of 10 ms, to catch blinks and other eye-related artifacts. Trial loss averaged 21.5% (Range: 0% - 76.7%). No individuals removed from analyses due to low trial numbers, given that the comparisons of interest were all within-subject. Artifact-free ERPs were averaged by stimulus type after subtraction of the 100 ms pre-stimulus baseline. Prior to statistical analyses, ERPs were digitally filtered with a low-pass filter of 30Hz.

## 6.3. Results

### *6.3.1. Survey Responses*

Since we were interested in individual differences in this study rather than group averages, we did not restrict participation based on handedness. Two of the 40 participants were left handed and 38 were right handed.

In the LEAP-Q, two participants reported that Spanish was their dominant language and 38 participants reported that English was their dominant language. Twenty-nine participants reported that they acquired English as their first language. One participant reported that French was their first language, three participants reported that German was their first language, four participants reported that Spanish was their first language, and three participants reported that another language (not English, Spanish, French or German) was their first language.

#### 6.3.1.1. English Proficiency

Twenty-nine of the participants reported that they began acquiring English starting in infancy. The remaining participants reported that they started learning English in childhood, listing the age at which they began learning English as being between 3 and 13 (mean = 8).

All of the participants reported that their proficiency in speaking, reading, and understanding spoken English was “very good” or better. When their proficiency ratings were converted to a 0-10 scale, their average proficiency rating were 9.3 for speaking, 9.5 for understanding spoken language, and 9.5 for reading.

#### 6.3.1.2. Spanish Proficiency

Twenty of the participants reported some knowledge of Spanish. Four people reported that it was their first language and that they had acquired Spanish before beginning to learn English. Another five participants reported that they had acquired both Spanish and English beginning in infancy. Two participants reported that they began learning Spanish in childhood, and the rest of the participants began to learn it at age 10 or later.

For the purposes of the modeling work described below, we considered a person to be proficient in a language if they rated themselves as being "very good" or better at speaking, understanding, or reading the language in question. For Spanish, there were 10 participants who met this criterion. When their survey responses were converted to the 0-10 numeric scale, these participants averaged 9.1 for speaking, 9.2 for understanding, and 8.7 for reading proficiency in Spanish.

There were also 10 less proficient participants, who gave themselves average ratings of 2.8 for speaking Spanish (range = 1-6), 2.6 for understanding spoken Spanish (range = 0-6) and 2.8 for reading Spanish (range = 1-6).

#### 6.3.1.3. French Proficiency

Thirteen of the participants reported some knowledge of French. Two people reported that it was their native language one other person reported that they began to learn French at age 2. The rest of the participants began to learn French at age 10 or later, typically as teenagers.

There were five participants who met our criteria for proficiency in French. These participants had average proficiency ratings of 8.2 for speaking French (range = 7-9), 8.0 for understanding spoken French (range = 6-9) and 8.0 for reading French (range = 7-9).

The eight less proficient participants gave themselves average ratings of 2.1 for speaking French (range = 1-5), 2.3 for understanding spoken French (range = 0-6) and 2.4 for reading French (range = 0-6).

#### 6.3.1.4. German Proficiency

Three participants reported that German was their first language and that they had started learning it in infancy. Another six participants reported some knowledge of German. One of those participants reported that they started learning German at age 4. The rest started learning German at age 10 or later, typically as teenagers.

Four participants met our criteria for proficiency in German. These participants gave themselves average ratings of 7.8 for speaking German (range = 7-8), 8.5 for understanding spoken German (range = 8-9), and 6.8 for reading German (range = 5-8).

The five less proficient participants gave themselves average ratings of 4.0 for speaking German (range = 1-7), 5.0 for understanding spoken German (range = 2-7) and 3.8 for reading German (range = 1-7).

### *6.3.2. Traditional ERP Analyses*

The EEG data was cleaned using independent components analysis (ICA) and the artifact rejections steps described in detail in Appendix A. To calculate the ERPs, bins were created based on every condition (i.e., English-Spanish translations) and every word as it appeared in every condition. Trials in those bins were time-locked to the onset of the stimulus and averaged together to create the ERPs.

The ERPLAB software was used to plot the grand average ERPs for all of the same-language word pairs to determine whether our dataset showed traditional N400 repetition effects in each language. In each case, we compared the same-language repetitions (i.e., English-English repetition) to the same-language unrelated pairs (i.e., English-English unrelated). The grand average ERPs for electrode Cz are shown in Figure 14 and the scalp maps for each language, with data averaged from 350-450 ms, are shown in Figure 15.

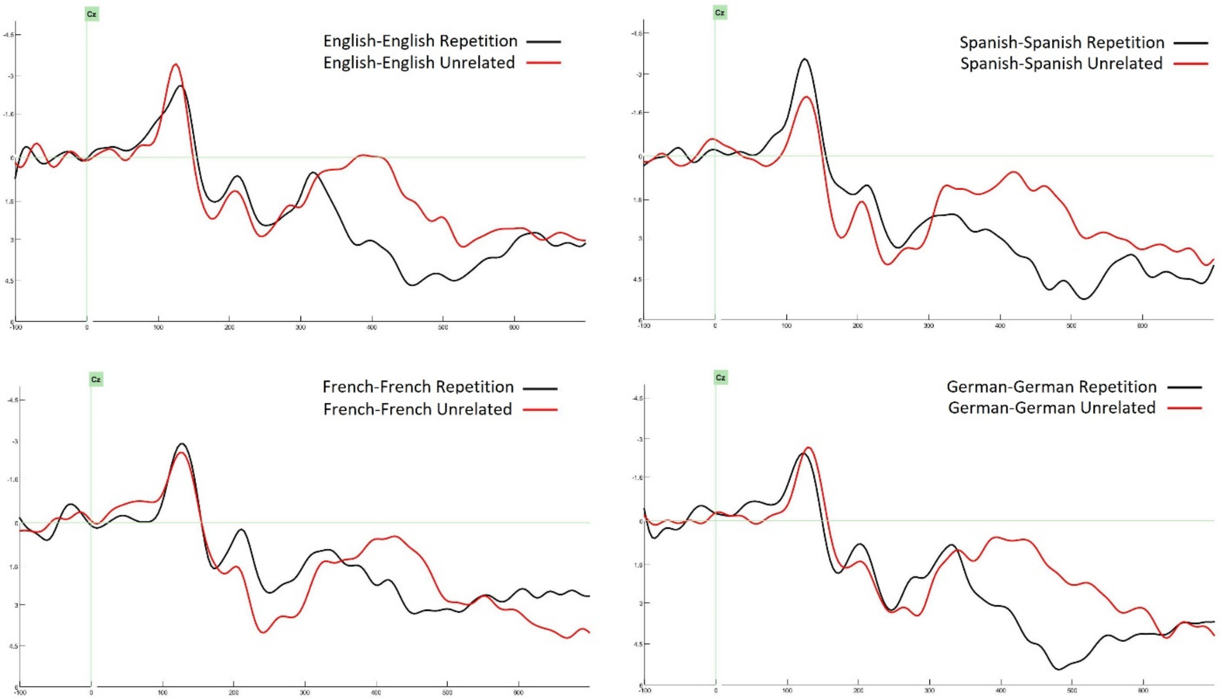

**Figure 14. Grand average ERPs from channel Cz for the same-language repetition and unrelated pairs in English, Spanish, French and German.**

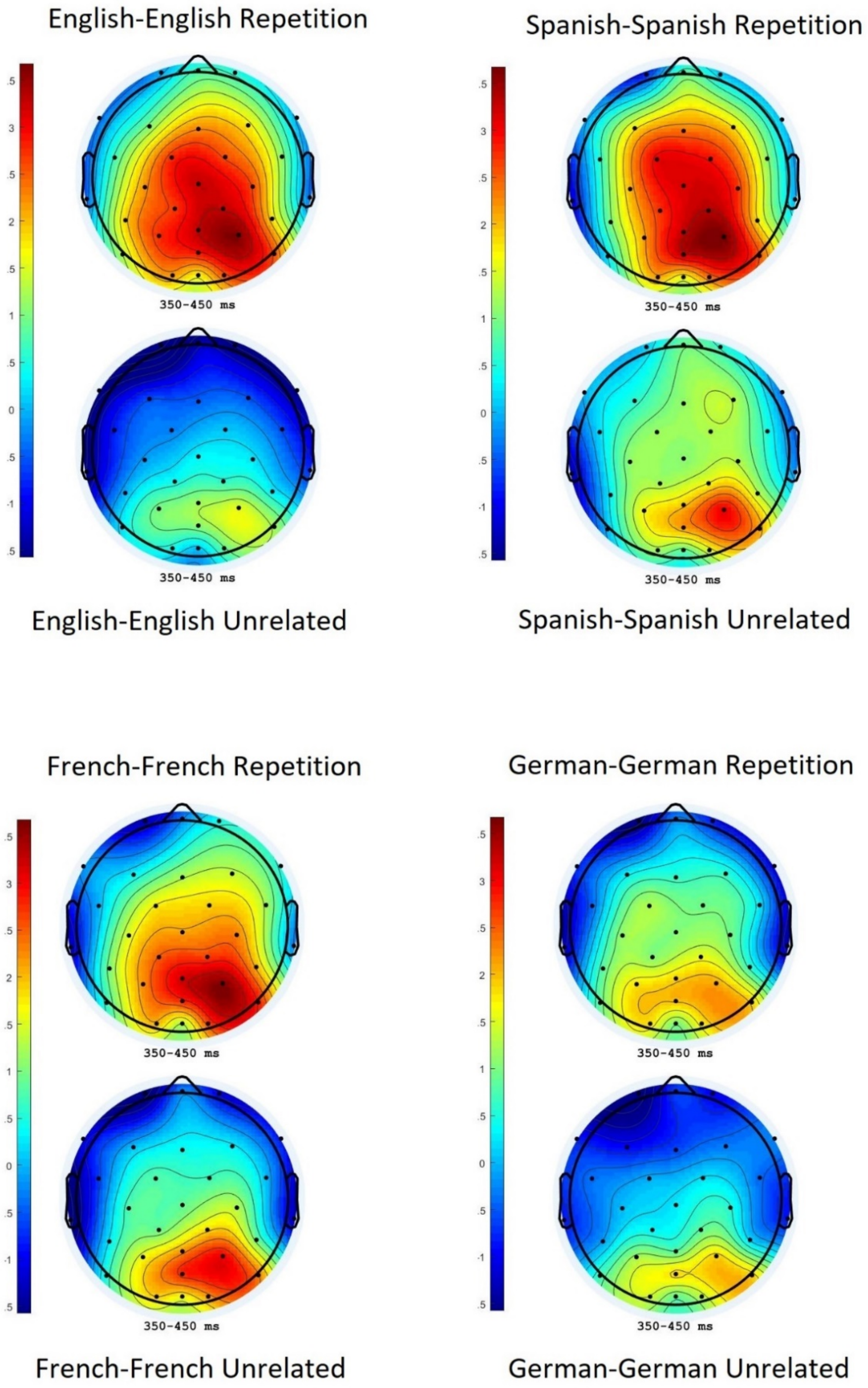

**Figure 15. Scalp maps showing the average voltage from 350-450 ms for repeated word pairs and unrelated word pairs in each language.**

These results illustrate that, on average, we observed N400 repetition effects in all four languages. Despite taking an unconventional approach of intermixing stimuli from different languages and intermixing participants with a variety of language backgrounds, these effects show up clearly in the ERPs for every language.

However, our primary interest lies in looking at individuals and the relationship between their language proficiency and their patters of ERPs in response to different conditions. As a preliminary analysis, compared four conditions: English-English repetition, Spanish-English translation, French-English translation, and German-English translation. For these four bins, the target words were always the same (the 30 English words in our stimulus set). In addition, the targets were always part of a semantically related pair. The only difference between the conditions was the language in which the prime word was presented. We predicted that if participants understood the language in which the prime words were presented, we would see an N400 repetition effect to the targets in that condition. If the participant did not understand the language in which the prime words were presented, we would not expect to see an N400 repetition effect for the target words.

For each participant, we generated scalp maps for these four conditions with the data averaged from 350-450 ms after stimulus onset. As expected, for many participants, a visual comparison of the scalp maps for these four conditions clearly indicated which languages they understood. Examples for three participants are shown in Figure 16. The first participant is a monolingual English speaker and showed an N400 repetition effect only in the English-English repetition condition. The second participant is proficient in both English and Spanish. They showed N400 repetition effects in English and Spanish, but not French or German. The third participant is proficient in English and French. This participant showed N400 repetition effects in English and French, but not Spanish or German.

This comparison was extremely simple, and the differences were not as visually apparent for all participants. However, this indicates that with carefully designed stimuli, N400 repetition effects can be used on the level of individual participants to make predictions about language proficiency.

In future work we will continue to analyze this dataset using a variety of techniques. We will perform more quantitative analyses of the N400 repetition effects for each participant and assess how many of our participants produced stable N400 repetition effects that accurately reflect their reported language proficiency. We will also explore how many trials are required for each participant to achieve stable effects. Finally, we will apply the ERP decoding technique (cf. Bae & Luck, 2019; Blankerts, Lemm, Treder, Haufe & Müller, 2011) to this dataset to explore its utility as another method for assessing language proficiency based on ERPs.

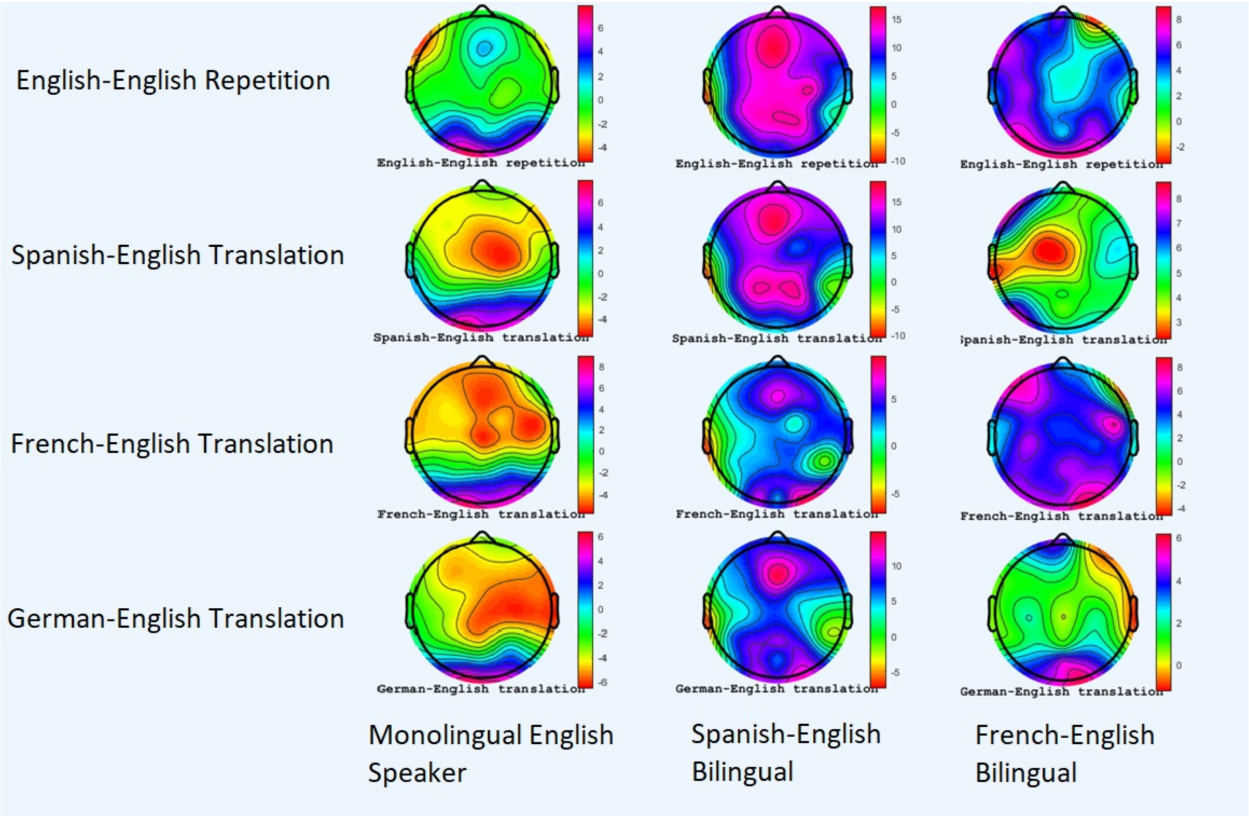
English-English Repetition
Spanish-English Translation
French-English Translation
German-English Translation
English-English repetition
Spanish-English translation
French-English translation
German-English translation
Monolingual English Speaker
Spanish-English Bilingual
French-English Bilingual

**Figure 16. Examples of scalp maps from three participants with different language backgrounds, showing the average voltage across the scalp from 350-450 ms.**

### *6.3.3. ERP Modeling Results*

As a preliminary pass at modeling the ERP data, we used a simple implementation of a method that was previously developed for ERP-based biometrics (Armstrong, Ruiz-Blondet, Khalifian, Kurtz, Jin & Laszlo, 2015). Given the relatively low numbers of French and German speakers in the dataset, we focused on predicting which participants were proficient in Spanish and which were not, an approach that mirrored our work with response time data in the behavioral experiments. Using only one channel (Cz), the maximum absolute value cross-correlation was computed between the participants averaged ERPs in the following conditions: Spanish-Spanish repetition, Spanish-Spanish unrelated, Spanish-English translation, Spanish-English unrelated, English-Spanish translation, English-Spanish unrelated. As a result, each participant had six ranked lists, one for each condition. Each list had 39 terms, one for each cross-correlation calculation (we removed the comparison of each participant with themselves).

Next, we computed the average precision for each participant's list, first assuming that the participant is not proficient in Spanish, and then assuming that they are proficient in both English and Spanish. To use terminology from information retrieval studies, an entry on each list was counted as "relevant" if the participant in question and the participant on that entry of their list have the same language proficiency. This resulted in a score for each assumption (non-proficient in Spanish or proficient in Spanish) for each condition (e.g., Spanish-Spanish repetition). For each condition, we predicted the proficiency of the individual to be the higher of the two average precision scores. For each participant, we predicted the proficiency of the individual based on the majority vote for all conditions.

The result of this simple first-pass analysis was a prediction accuracy of 75% correct. This exceeded the accuracy of any of the models that were based on behavioral data alone, supporting our hypothesis that the rich data provided by ERPs is more predictive of language proficiency than behavioral data alone. In future work, we will continue to develop the model of ERP data, using more sophisticated modeling techniques, in an effort to further improve the model's predictive power.

# 7. GENERAL DISCUSSION

Overall, this project demonstrated that there is promise in the idea of using machine learning methods to assess individual differences between participants. In both behavioral and EEG experiments, we identified tasks that allowed us to make predictions about individual participants' proficiency in different languages. Although not all of the tasks we tested were successful and our prediction accuracy has not yet exceeded 75% correct, we have identified promising avenues for future research.

We chose to focus on language processing in this project because of the well-characterized ERP effects related to language processing. When our plans for collection EEG data were postponed due to the COVID-19 pandemic, we turned to online remote data collection, which necessitated the development of behavioral tasks that participants could complete without understanding all of the languages represented in the stimuli. Using word length judgement tasks, we were able to make reasonably good predictions about which participants were proficient in Spanish and which were not. Interestingly, semantic priming did not appear to be the driving factor behind the model's predictions. Across all of our behavioral experiments, the model performed better when using response time data from every trial rather than priming effects.

We also explored using Stroop tasks to see if modeling interference effects would allow for more accurate predictions of language proficiency. However, our online implementations of the Stroop task on Amazon Mechanical Turk produced data that were too noisy to be useful for modeling purposes. Furthermore, the proficient Spanish speakers in these tasks did not show Stroop effects for the Spanish stimuli. It is not clear whether this pattern was due to the implementation of the task or to the bilingual advantage in resistance to semantic interference effects that has been observed in some prior studies. In either case, these data were not useful for making predictions about which participants were proficient in which languages.

Finally, at the end of this project we were able to collect EEG data for the task we had originally planned to use for this project. Our preliminary analyses show that N400 repetition effects can be used to make predictions about individual participants' language proficiency. Initial modeling results suggest that predictions based on ERPs will be more accurate than predictions based on behavioral responses alone. In a future project, we will conduct additional analysis of our ERP data from this experiment and further develop our approach to using machine learning methods to model and make predictions about individual differences in human cognitive performance. As we continue to make progress in this area, we aim to bring the large literature on individual differences in cognition to bear on specific, applied problems. Our continuing work will develop new methods that will allow us to make these connections, supporting improvements in human performance for specific individuals in specific settings.

# REFERENCES

[1] Badzakova-Trajkov, G., Barnett, K.J., Waldie, K.E. and Kirk, I.J., 2009. An ERP investigation of the Stroop task: The role of the cingulate in attentional allocation and conflict resolution. Brain research, 1253, pp.139-148.

[2] Bae, G. Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, *184*, 242-255.

[3] Bialystok, E., 1999. Cognitive complexity and attentional control in the bilingual mind. Child development, 70(3), pp.636-644.

[4] Bialystok, E., Craik, F. and Luk, G., 2008. Cognitive control and lexical access in younger and older bilinguals. Journal of Experimental Psychology: Learning, memory, and cognition, 34(4), p.859.

[5] Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K. R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, *56*(2), 814-825.

[6] Chen, H.C. and Ho, C., 1986. Development of Stroop interference in Chinese-English bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12(3), p.397.

[7] Coderre, E. L., & Van Heuven, W. J. (2014a). Electrophysiological explorations of the bilingual advantage: Evidence from a Stroop task. PloS One, 9(7).

[8] Coderre, E. L., & Van Heuven, W. J. (2014b). The effect of script similarity on executive control in bilinguals. Frontiers in Psychology, 5(1070), 1–16.

[9] Coderre, E.L., Van Heuven, W.J. and Conklin, K., 2013. The timing and magnitude of Stroop interference and facilitation in monolinguals and bilinguals. Bilingualism: Language and Cognition, 16(2), pp.420-441.

[10]Costa, A., Hernández, M. and Sebastián-Gallés, N., 2008. Bilingualism aids conflict resolution: Evidence from the ANT task. Cognition, 106(1), pp.59-86.

[11]De Bruin, A., Treccani, B. and Della Sala, S., 2015. Cognitive advantage in bilingualism: An example of publication bias?. Psychological science, 26(1), pp.99-107.

[12]Delorme, A. & Makeig, S. (2004) EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics, *Journal of Neuroscience Methods 134*, 9-21.

[13]Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808-816.

[14]Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, *30*, 149-166.

[15]Fang, S.P., Tzeng, O.J. and Alva, L., 1981. Intralanguage vs. interlanguage Stroop effects in two types of writing systems. Memory & Cognition, 9(6), pp.609-617.

[16]Goldfarb, L. and Tzelgov, J., 2007. The cause of the within-language Stroop superiority effect and its implications. Quarterly Journal of Experimental Psychology, 60(2), pp.179-185.

[17]Haass, M. J. & Matzen, L. E. (2011). Using computational modeling to assess use of cognitive strategies. *Foundations of Augmented Cognition: Directing the Future of Adaptive Systems. Lecture Notes in Artificial Intelligence, 6780/2011*, 77-86.

[18]Haass, M. J., Matzen, L. E., Butler, K. M., & Armenta, M. (2016, March). A new method for categorizing scanpaths from eye tracking data. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 35-38). ACM.

[19]Heidlmayr, K., Hemforth, B., Moutier, S., & Isel, F. (2015). Neurodynamics of executive control processes in bilinguals: Evidence from ERP and source reconstruction analyses. Frontiers in Psychology, 6(821), 1–17.

[20]Heidlmayr, K., Moutier, S., Hemforth, B., Courtin, C.,Tanzmeister, R., & Isel, F. (2014). Successive bilingualism and executive functions: The effect of second language use on inhibitory control in a behavioural Stroop Colour Word task. Bilingualism: Language and Cognition, 17(03).

[21]Hernandez, A.E. and Li, P., 2007. Age of acquisition: its neural and computational mechanisms. Psychological bulletin, 133(4), p.638.

[22]Hilchey, M.D. and Klein, R.M., 2011. Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. Psychonomic bulletin & review, 18(4), pp.625-658.

[23]Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology, 82*(3), 426-433.

[24]Jankowiak, K., & Rataj, K. (2017). The N400 as a window into lexico-semantic processing in bilingualism. *Poznan Studies in Contemporary Linguistics*, *53*(1), 119-156.

[25]Khalifian, N., Stites, M. C., & Laszlo, S. (2015). Relationships between event-related potentials and behavioral and scholastic measures of reading ability: A large-scale, cross-seciontal study. *Developmental Science, 19*(5), 723-740.

[26]Kiefer, M. (2002). The N400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of N400 priming effects. *Cognitive Brain Research*, *13*(1), 27-39.

[27]Kiefer, M., & Brendel, D. (2006). Attentional modulation of unconscious "automatic" processes: Evidence from event-related potentials in a masked priming paradigm. *Journal of Cognitive Neuroscience*, *18*(2), 184-198.

[28]Kousaie, S. and Phillips, N.A., 2012a. Ageing and bilingualism: Absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. Quarterly Journal of Experimental Psychology, 65(2), pp.356-369.

[29]Kousaie, S. and Phillips, N.A., 2012b. Conflict monitoring and resolution: Are two languages better than one? Evidence from reaction time and event-related brain potentials. Brain research, 1446, pp.71-90.

[30] Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.

[31]Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207* (4427), 203-205.

[32]Kutas, M., & Van Pettern, C. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. San Diego, CA, US: Academic Press.

[33]Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., & Kliegl, R. (2018). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*, 1-18.

[34]Lee, T.M. and Chan, C.C., 2000. Stroop interference in Chinese and English. Journal of Clinical and Experimental Neuropsychology, 22(4), pp.465-471.

[35]Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213.

[36]Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature, 383*(6601), 616-618.

[37]McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature neuroscience*, *7*(7), 703.

[38]McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*, 123-150.

[39] Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish–English bilinguals. *Cognitive Brain Research*, *22*(2), 205-220.

[40]Moreno, E. M., Rodriguez-Fornells, A., & Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *Journal of Linguistics, 21*, 477-508.

[41]Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory & cognition*, *33*(6), 972-983.

[42]Okada, K., He, G. and Gonzales, A., 2019. Monolinguals and Bilinguals Differ in Performance on the Taboo Stroop Task. The Open Psychology Journal, 12(1).

[43]Osterhout, L., Poliakov, A., Inoue, K., McLaughlin, J., Valentine, G., Pitkanen, I., Frenck-Mestre, C., & Hirschensohn, J. (2008). Second-language learning and changes in the brain. *Journal of neurolinguistics*, *21*(6), 509-521.

[44]Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, *62*(8), 1457-1506.

[45]Rolke, B., Heil, M., Streb, J., & Hennighausen, E. (2001). Missed prime words within the attentional blink evoke an N400 semantic priming effect. *Psychophysiology*, *38*(2), 165-174.

[46]Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist*, *40*(3), 355.

[47]Stites, M. C., Federmeier, K. D., & Christianson, K. (2016). Do morphemes matter when reading compound words with transposed letters? Evidence from eye-tracking and event-related potentials. *Language, Cognition and Neuroscience, 31*(10), 1299-1319.

[48]Stites, M. C., & Laszlo, S. (2015). How do random effects strucutres impact LMER outcomes in an ERP study? *Psychophysiology, 52*, S116-S116.

[49]Stites, M. C., & Laszlo, S. (2017). Time will tell: A longitudinal investigation of brain-behavior relationships during reading development. *Psychophysiology, 54*(6), 798-808.

[50]Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150* (3700), 1187-1188.

[51]Tanner, D., Goldshtein, M., & Weissman, B. (2018). Individual differences in the real-time neural dynamics of language comprehension. In *Psychology of learning and motivation* (Vol. 68, pp. 299-335). Academic Press Cambridge, MA.

[52]Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological bulletin*, *135*(3), 452.

[53]Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: an electric sign of sensori-motor association and expectancy in the human brain. *Nature*, *230*, 380-384.

[54]Weber, K., & Lavric, A. (2008). Syntactic anomaly elicits a lexico-semantic (N400) ERP effect in the second language but not the first. *Psychophysiology*, *45*(6), 920-925.

[55]Okuniewska, H., 2007. Impact of second language proficiency on the bilingual Polish-English Stroop task. Psychology of language and Communication, 11(2), pp.49-63.

[56]MacLeod, C.M., 1991. Half a century of research on the Stroop effect: an integrative review. Psychological bulletin, 109(2), p.163.

[57]Mägiste, E., 1984. Stroop tasks and dichotic translation: The development of interference patterns in bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(2), p.304.

[58]Mägiste, E., 1985. Development of intra-and interlingual interference in bilinguals. Journal of Psycholinguistic Research, 14(2), pp.137-154.

[59]Mägiste, E., 1992. Second language learning in elementary and high school students. European Journal of Cognitive Psychology, 4(4), pp.355-365.

[60]Paap, K.R. and Greenberg, Z.I., 2013. There is no coherent evidence for a bilingual advantage in executive processing. Cognitive psychology, 66(2), pp.232-258.

[61]Paap, K.R., Johnson, H.A. and Sawi, O., 2016. Should the search for bilingual advantages in executive functioning continue?.

[62]Sabourin, L., Vinerte, S. and Mayo, M.D.P.G., 2015. The bilingual advantage in the Stroop task: simultaneous vs. early bilinguals. Bilingualism, 18(2), p.350.

[63]Singh, N. and Mishra, R.K., 2012. Does language proficiency modulate oculomotor control? Evidence from Hindi-English bilinguals. Bilingualism, 15(4), p.771.

[64]Stroop, J.R., 1935. Studies of interference in serial verbal reactions. Journal of experimental psychology, 18(6), p.643.

[65]Sumiya, H. and Healy, A.F., 2004. Phonology in the bilingual Stroop effect. Memory & cognition, 32(5), pp.752-758.

[66]Sutton, T.M., Altarriba, J., Gianico, J.L. and Basnight-Brown, D.M., 2007. The automatic access of emotion: Emotional Stroop effects in Spanish–English bilingual speakers. Cognition and Emotion, 21(5), pp.1077-1090.

[67]Tse, C.S. and Altarriba, J., 2012. The effects of first-and second-language proficiency on conflict resolution and goal maintenance in bilinguals: Evidence from reaction time distributional analyses in a Stroop task. Bilingualism, 15(3), p.663.

[68]Van den Noort, M., Vermeire, K., Bosch, P., Staudte, H., Krajenbrink, T., Jaswetz, L., Struys, E., Yeo, S., Barisch, P., Perriard, B. and Lee, S.H., 2019. A systematic review on the possible

relationship between bilingualism, cognitive decline, and the onset of dementia. Behavioral Sciences, 9(7), p.81.

[69]Yow, W.Q. and Li, X., 2015. Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: why variations in bilingual experiences matter. Frontiers in psychology, 6, p.164.

[70]Woumans, E., Ceuleers, E., Van der Linden, L., Szmalec, A. and Duyck, W., 2015. Verbal and nonverbal cognitive control in bilinguals and interpreters. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5), p.1579.

[71]Zied, K.M., Phillipe, A., Karine, P., Valerie, H.T., Ghislaine, A. and Arnaud, R., 2004. Bilingualism and adult differences in inhibitory mechanisms: Evidence from a bilingual Stroop task. Brain and cognition, 54(3), pp.254-256.

[72]Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152).

[73]Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In 2010 *20th International Conference on Pattern Recognition* (pp. 3121-3124). IEEE.

[74]Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

[75]Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

[76]Dijkstra, A. F. J., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175-197.

[77]Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.

[78]Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.

[79]Grainger, J., & Frenck-Mestre, C. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and Cognitive Processes, 13*(6), 601-623.

[80]Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287-317.

[81]James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

[82]Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149-174.

[83]Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism*, *13*(3), 373.

[84]Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE 7(8)*: e43230.

[85]Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50(4)*, 940-967.

[86]Martin, C. D., Dering, B., Thomas, E. M., & Thierry, G. (2009). Brain potentials reveal semantic priming in both the 'active' and the 'non-attended' language of early bilinguals. *NeuroImage*, *47*(1), 326-333.

[87]Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

[88]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

[89]Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

[90]Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

[91]Real Academia Española (2010). *Ortografía de la lengua Española*. Espasa.

[92]Schoonbaert, S., Duyck, W., Brysbaert, M., & Hartsuiker, R. J. (2009). Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & Cognition*, *37*(5), 569-586.

[93]Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

[94]Van Hell, J. G., & Tanner, D. (2012). Second language proficiency and cross-language lexical activation. *Language Learning*, *62*, 148-171.

# APPENDIX A.

**Identify the correct programs and plug-ins to install to run MATLAB, EEGLAB, and ERPLAB for data processing**

1. MATLAB
    a. Request in Nile: https://nile.sandia.gov/services/1370
    b. Instructions for installing:
        i. https://wiki.sandia.gov/display/CEE9/SRN+-+MATLAB+and+Simulink+Install+and+Launch
2. EEGLAB
    a. https://sccn.ucsd.edu/eeglab/downloadtoolbox.php
    b. put EEGLAB in your MATLAB path
    c. See wiki for help:
        i. https://eeglab.org/tutorials/01_Install/Install.html
    d. Download ANT extension to correctly load data: https://sccn.ucsd.edu/eeglab/plugin_uploader/plugin_list_all.php
3. ERPLAB
    a. https://erpinfo.org/erplab
    b. put unzipped ERPLAB folder in the EEGLAB 'plugins' folder
    c. Make sure that you see "EEGLAB: adding "erplab" v8.20 (see >> help eegplugin_erplab)" when loading EEGLAB to ensure that ERPLAB has been loaded
    d. https://github.com/lucklab/erplab/wiki
4. Test data sets to play around with:
    a. CORE: https://osf.io/thsqg/files/
    b. N400 from CORE: https://github.com/lucklab/ERP_CORE
5. Download and install binica, which is supposed to be much faster to run than runica because it is a compiled version of the same script
    a. https://github.com/lucklab/lucklab_installBinica
    b. **Note**: I had a lot of trouble getting this to run. I had to edit the install_binica script to make sure that my folder containing eeglab was actually in the path
    c. I also had to change an additional line of the icadefs.m script so that all of the .m files were being searched for in the right folder

## Processing Steps Overview

- There are a few steps that need to be done manually, but much of the work can be carried out in a batch. This is helpful if you want to set the artifact rejection parameters for people automatically, but then you need to re-run the processing steps to change the bdf, apply a different filter, etc
- The data must be loaded from a cnt file manually.
- Then, you can run Script 1 to do initial processing (adding channel locations, rereferencing, high pass filtering, etc.)
- If you ARE doing ICA correction, run scripts 2-4 as described next. If not, skip scripts 2-4 and go right to script 5
    - Run scripts 2 & 3 in a row.
    - Stop, look at the output of 3, choose which components to reject, and add to the spreadsheet.
    - Run script 4
- Run script 5
- Perform artifact rejection for each person individually. Start with the baseline thresholds, change as necessary to catch artifacts, and then add the values that were used to the appropriate spreadsheets
- Run script 6 (artifact rejection), 7 (averaging), 8 (plot individual values)
- Update text file of people to include in grand averages, and then run script 9
- Update and run script 12, to measure ERP values for statistical analyses
- There are 3 different versions of the scripts:
    - _lang.m are the baseline scripts, edited from the Luck Lab materials
    - _lang_fcn.m are the same scripts as above, but executed as functions. This is helpful if want to run them all together and only want to update the SUB variable (containing subject numbers) once.
    - _lang_ica.m are the same as the baseline scripts, but they operate on the ica-corrected data instead. The main differences are:
        - 1. The ICA data is saved in a different folder
        - 2. The artifact rejection values should be different; namely, we would only want to reject blinks that occur during the stimulus presentation period rather than the whole trial
- General EEGLAB / ERPLAB tips:
    - All steps can be carried out through the GUI. However, for reproducibility and efficiency, we will script our analyses. It is very difficult to move between scripting and GUI, so it is best to pick one method and stick with it. This is why the start of each script clears the workspace and restarts EEGLAB frequently – it is really easy to get the two out of sync .
    - If you want to reproduce the steps you carried out in the GUI, you can type eegh into the command line, and it will produce a history of commands that have been run. You can *typically* copy/paste these into your script, although sometimes the set number being referred to is off. Do so with caution.

## Processing Steps Detail:

**Manual Steps:**

Opening lang data: this is unfortunately a manual process the first time you do it.

1. File –> import data –> using eeglab functions and plugins –> from ANT EEProbe CNT file
    a. This will pop up a file explorer window. Navigate to the subject's folder and click on the .cnt file
    b. I did not choose anything for the interval; this seems to import the entire file just fine
2. Plot – channel data (scroll)
    a. Make sure there is data there. You will likely need to make the scale very large because the data isn't filtered (like, 500+). I suggest clicking through the data to make sure that item codes appear, and that you seem to have ongoing data from all channels.
3. Save as:
    a. #.set (just the subject's number), in a folder that is also just the subject's number
    b. Click 'save as' from EEG lab, and it makes a set file
    c. For the sake of the processing script, be sure to save into the subject's folder (also just the subject number).
        i. Note: you could also save them in the main directory, you will just need to update the DIR where MATLAB looks for the data in the scripts
    d. This is a good time to make sure that the subject's folder also has a folder called "graphs" in order to save plots into later

Run script 1 to take care of very basic initial processing (high pass filter, rereference to average of mastoids, apply channel names, etc)

4. Script 1
    a. The channels are labeled with their names but not with their scalp locations. You need to have the scalp locations paired to make sure that plotting the ICA components works properly.
    b. There are 34 channels
        i. there's just 1 that says HEOG – it is bipolar
        ii. There's also just 1 that says VEOG – it is also bipolar
    c. I assume we are using the 32-channel waveguard cap. I wonder if there is a channel file for that automatically in ERP lab?
        i. https://www.ant-neuro.com/products/waveguard/electrode-layouts
        ii. I found one (in a link below), but it didn't load correctly. So, I just used the default suggested electrode layout for the 10-20 system, as Laura suggested. Because we are not doing source localization, I think this is okay.
    - For additional help: https://github.com/lucklab/erplab/wiki/EEG-and-ERP-Channel-Operations

ICA Prep and Application:

5. Script 2:
    a. We will use this to delete segments of EEG that are more than 2000 ms away from a code (i.e., if the recording was left on during a break and there is a lot of big noise, this will mess with the creation of ICA components). I will NOT be following the Luck recommendation of also visually determining "especially noisy" epochs to delete. I have commented these sections out. They can be added back in later if desired.
    b. https://github.com/lucklab/erplab/wiki/Continuous-EEG-Preprocessing#delete-time-segments
    c. This article helped me figure out what the different parameter settings mean
6. Script 3:
    a. Decompose using ICA
    b. Need to install the binica script
        i. https://github.com/lucklab/lucklab_installBinica
        ii. Follow the read-me instructions! I had to make a few of my own edits in addition to those in the script
            1. Path: I had to add eeglab to the path because otherwise the path command would not work
            2. There were 2-3 lines where I had to change the icadefs.m file…all within the same if/else statement though
    c. After this script runs, you need to:
        i. Look at the ICA components that were saved into the plot for each person
        ii. Update the excel sheet "ICA_Components_N400_lang.xlsx" with the component(s) to be removed for each person
        iii. In the N400 CORE dataset, steve luck has an example of 40 subjects, their ICA component scalp maps, and which ones they deleted, for reference
7. Script 4: remove ICA components
    a. This seemed to run without issue
    b. Note that if you decide to remove other ICA components, you can just change the number in the spreadsheet and re-run this script

Pick back up here if you are not doing ICA correction.

8. Script 5
    a. Applies the bdf file listed in the script to create bins
    b. I made 2 bdf files: one for the main condition-level bins and one for each word each time it appeared in a different bin
    c. I did not do any time-locking to the responses OR to the prime words
    d. Note that you can make as many bins as you would like! More bins will generally take longer at this stage and at averaging. You can also re-create the bdf file if you want to change bins, and re-run this stage forward.

Stop. You will carry out the steps listed in the section using the GUI, add the threshold values to a spreadsheet, and THEN run the script to process the data.

9. Script 6: artifact rejection
    a. This is the most difficult step to do in an automated and repeatable way. The best solution I can come up with, based on recommendations from Steve Luck's materials is:

i. Run all scripts UP TO script 5. Then, run each of the artifact rejection categories individually. Start by applying the baseline parameters, view the scroll data to see which trials are marked for rejection for that subject.
ii. If it looks like that parameter setting does a reasonable job of rejecting trials it should and not rejecting trials it shouldn't, and then add those to the excel file for each subject. THEN, use the parameters set for each subject to delete the artifacts in an automated way using the script.
iii. We will **not** do hand-selection of trials to reject: this is too subjective, time-consuming, and impossible to describe/repeat. They all must fall within a filter's parameters.

b. I am only using 2 artifact rejection steps to balance the time it takes to go through every subject's data so many times and the diminishing returns of running each step. Steve Luck's lab suggests having potentially up to 4-5 different steps.
c. I applied a low pass filter at 30 Hz before artifact rejection. Instructions for that: EEGLAB –> ERPLAB –> Filter & Frequency tools –> Filters for EEG data -> Apply. Save the set file to the workspace, then your data during artifact rejection will look like that data that you are going to be working with later to average, etc.
d. Pay attention to the scale of the data, and set the same scale for all subjects to maintain consistency. I work with a scale around 90. The smaller you go, the more junk you see. The more you zoom out, the better/flatter your data looks. Either way can be deceptive if you're not used to one or the other.
e. It can sometimes be helpful to look through several subjects worth of data before jumping into artifact rejecting a single participant. This can help get a good idea of what "good" data looks like, what typical artifacts are, etc.
f. The two artifact rejection steps I used:
   i. Identify CRAP (commonly recorded artifactual potentials) using simple voltage threshold algorithm
      1. How to set in EEGLAB:
         a. In EEGLAB window, click: ERPLAB -> Artifact Rejection in epoched data -> simple voltage threshold
         b. Will apply this filter this for **all channels except** HEOG and VEOG (available channels – click browse, and then select all that are not HEOG and VEOG) [1:30]
         c. In the selection window, you can see what the parameters are for marking to reject (default I think it is -200 200)
      2. In the window that pops up, epochs that are marked with yellow are marked to reject based on these thresholds. Your main task is to scroll through these marks, see if they look reasonable (is it rejecting most trials that seem like they should be rejected? And not rejecting trials that look okay?). If so, record these thresholds in the spreadsheet. do not click/unclick to manually reject or not. The offending channels will be highlighted in red.
      3. If it seems like there are trials that should/not be rejected, close the window, don't save the dataset, and re-run with a new threshold to see how that changes things. If it's rejecting too much, increase the numbers so that the filter catches less. If it's not rejecting enough,

lower the numbers. Est: move in steps of ~20-30 to see how that changes things.

4. Note: we will always reject the WHOLE TRIAL, will never drop single channels.
5. These values get saved into the spreadsheet: AR_1_Parameters_for_SVT_CRAP_lang.xlsx
    a. Note that in the column Channels, if it says "default", that will get changed in script 6 to be channels 1:30. There is just no good way to include this in excel and have it be read into matlab correctly.
    b. If you want to set a different default set of channels (i.e., if you ran with more or fewer electrodes, just change this parameter in the script! You can make as many as you want/need).
6. Close the window, and do NOT have ERPLAB save a new set file.

ii. Identify blinks in VEOG using moving window peak-to-peak threshold

1. How to set up in EEGLAB:
    a. In EEGLAB window, click: ERPLAB -> Artifact Rejection in epoched data -> moving window peak-to-peak threshold
    b. Will apply this filter this for **ONLY** VEOG (available channels –select VEOG) [32].
    c. Voltage threshold: 180
        i. You can make this bigger or smaller if necessary
    d. Moving window full width: 175
    e. Window step: 10
    f. Test period: -100 900
        i. If you are running this step on ICA-corrected data with blink artifacts removed, you may want to set the time window to only detect blinks during the stimulus presentation window. This would help ensure that trials for which the person did not see the stimulus are removed.
    g. Note: I did not also do a separate step for eye movements, as those tended to be picked up by the VEOG as well. For some people who had super large eye movements, I just added the HEOG Channel (32) to this step. This is not a great way to do it! We would rather have those trials detected in separate steps.
2. Same as above, a scroll window will pop up and show you which trials were rejected based on this filter. It will not show you which trials were rejected based on the previously applied filters. If there are big artifacts, multiple filters will probably catch them. That is okay for our purposes. The important thing is to look at all trials to make sure that each filter deletes the trials it should, but does not unnecessarily delete a lot of extra trials.
3. When you are done, save the values into the spreadsheet AR_3_Parameters_for_MW_Blinks_lang.xlsx. Note that for the

channels, if it says "both" I have created a variable in the script 6 that changes this to 31 and 32. Again, you can alter this variable if you'd like. Otherwise, just put the channel number that you used for the analysis.

10. Script 7: averaging
    a. This step is pretty straightforward—it will create an average for every bin in the dataset. Trials can belong to multiple bins, so that is not an issue.
    b. It will also create a spreadsheet that describes the number of trials accepted/rejected per bin, and the overall percentage of trials rejected for that individual. This is nice when reporting the overall values in your methods section. It can also be a good way to know which people need to be removed from analyses (i.e., if more than 25% of trials are rejected).
    c. In this script, I have commented out a section to create difference waves. Difference waves are a great way to compare two conditions directly when all you are interested in is the *difference* between conditions/bins rather than the overall morphology of the waveform. It references a text file that is saved in the workspace to allow the creation of difference waves.
11. Script 8: plotting
    a. This script is nice because it will create a plot for every person and save it in their folder. I'm having a hard time getting MATLAB to actually save the erp plots in a nice-looking way, but it is a good practice to look at each person's averages before you make grand averages to make sure that there aren't any huge obvious artifacts that you've rejected
    b. It can help at this stage if you've made a bin of all trials; that way, you can have one ERP that contains all of the non-rejected trials to see what they look like
    c. Be sure to make the 'graphs' folder before you try to save from the script, otherwise it will throw you an error.
12. Script 12: measuring ERPs (merp)
    a. The measurements we tend to do our stats on are: mean amplitude from 200-400 ms at every electrode for each condition for each person—and then we will run ANOVAs on those.
    b. Use caution when measuring peak amplitude; this can differ dramatically based on filtering, time window, etc., and is not always most informative.
    c. I created a different merp script for the ML-like analyses, with "allwords" in the title. It will create 256 output text files (one for every sample), taking a measurement once every 3.906 ms from -100 to 900 ms, at Cz, for every bin and every person.

# DISTRIBUTION

### Email—Internal

| Name | Org. | Sandia Email Address |
|---|---|---|
| Susan Adams | 06672 | smsteve@sandia.gov |
| Technical Library | 01977 | sanddocs@sandia.gov |

Hardcopy—Internal

| Number of Copies | Name | Org. | Mailstop |
|---|---|---|---|
| 1 | Laura Matzen | 06672 | 1327 |

This page left blank

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.