Effects of bilingualism on inhibition unlikely- Evidence from a Bayesian Inquiry

Moulshree Rana (moulshree20@iitk.ac.in)
Department of Cognitive Science, IIT-Kanpur,
Kanpur - 208016 India

Ark Verma (arkverma@iitk.ac.in)
Department of Cognitive Science, IIT-Kanpur,
Kanpur - 208016 India

Abstract

Recent literature indicates that the effects of bilingualism on executive control functions need to be examined with a more comprehensive characterization of bilingualism, and with the use of multiple measures of executive control (Backer and Bortfeld, 2021,Paap and Greenberg, 2013). In the current study, we operationalize bilingualism using a set of continuous variables related to language knowledge and use. We examine the effects of language proficiency, immersion, language dominance, diversity of language use and language switching on individuals' performance on tasks measuring inhibition. 66 Hindi-English bilinguals responded to the LHQ3, BSWQ and completed four inhibition tasks online. Inhibition tasks varied on the type of conflict (stimulus-stimulus/ stimulus-response) and the type of stimuli (arrows/ words). Bilingualism related variables failed to predict performance on any of the four tasks when included in linear regression models. We also conducted Bayesian regression analyses to validate the evidence for the lack of an effect. For three out of four tasks, we find BF_{10} (Bayes Factor indicating evidence for the alternate over null) of less than 1. Our data were most likely for the case where the null is true. Posterior odds for the null increased by factors of 13.1, 10.9, 4.3 and 11.8 for the four tasks, respectively. However, for the nonverbal Vertical Stroop Task, the best model contained only multilingual diversity scores as a predictor. We fail to find adequate evidence for the effects of bilingualism on inhibitory performance. We find that the effects of bilingualism do not appear to be task-specific or dependent on the type of conflict involved in a task, as previously suggested.(Blumenfeld and Marian, 2014) and conclude that it is unlikely that behavioral effects of bilingualism on inhibition exist.

Keywords: bilingual advantage; inhibition; bayesian; task-specificity

Introduction

Several studies across the past decade have participated in the debate regarding a "Bilingual Advantage" on executive functions, specifically inhibition ability (Bialystok et al., 2010,Bialystok, 2011,Paap and Greenberg, 2013,Antón et al., 2014). The notion of facilitatory (or other) effects of bilingualism is based on the following reasoning: When bilinguals wish to produce words in a target language, translational equivalents from the non-target language are also active and compete (Blumenfeld and Marian, 2014). The bilingual must then engage some conflict resolution mechanism in order to produce correct speech in the language of their choice without intrusions. Practice with such conflict resolution is thought to transfer to non-linguistic domain-general tasks which also involve inhibition of irrelevant representations. This simple reasoning needs to be qualified with other facts: Effects of bilingualism would only appear if the mechanism for resolving conflict between lexical representations is also used for resolving general conflict. In addition, the use of two languages and the practice with inhibition achieved must have effects that go beyond the practice achieved by resolving conflict within the use of a single language. (for example, in selecting between words depending on formal or informal contexts, or choosing between synonyms)

Mixed results

To demonstrate the effects of bilingualism on inhibition, studies typically compare the performance of groups of monolinguals and bilinguals on non-verbal inhibition tasks, such as the Stroop, Simon or Flanker tasks. However, such group comparisons have yielded mixed results: For instance, while some studies found that bilinguals outperformed their monolingual counterparts, several recent studies fail to find a non-trivial difference. This is demonstrated by several meta-analyses showing small effect sizes and results consistent with the absence of bilingual advantages. (Lehtonen et al., 2018,Paap, Mason, et al., 2020)

Mixed findings remain open to two exclusive interpretations. One, effects of bilingualism exist, and appear under certain conditions only and not others, or Two, bilingualism does not affect inhibition and any effects seen are artefacts. For the first, we would need to specify what are the conditions under which we should observe the advantage associated with bilingualism. In the second case, we would need to assess the hypothesis with stricter controls and robust theories. With reference to the initial reasoning provided (where repeated practice leads to a transfer), problems in examination of the bilingual advantage in EF hypothesis may arise either at source (due to the modular nature of bilingual language control), transfer (due to failure in transfer) or target (due to unreliable measurement of inhibition) domains. (Blanco-Elorrieta and Caramazza, 2021)

Operationalising Bilingualism

The examination of the effects of bilingualism as a dichotomous yes/no question, answered based on group comparisons, is not adequate. For instance, when a study reports differences in performance of groups of monolinguals and bilinguals, one cannot readily describe the aspects of bilingualism driving these differences. Studies examining group-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

based differences face criticisms citing "a forest of confounding variables" and the inadequacy of treating bilingualism as a unitary construct. It has therefore been suggested that patterns of language use might be what is important, rather than mere membership in a bilingual or monolingual group. Studies with this temperament have examined various sources of individual differences in language use, for instance, Age of acquisition of the L2, Proficiency in L2, frequency of language switching and immersion in various language contexts.(Zirnstein et al., 2019, Dash and Kar, 2020, Beatty-Martinez et al., 2020, Pot et al., 2018)

Comparison across the two groups is also susceptible to confounds, such as those of culture, since several studies feature immigrant bilinguals who are compared to culturally different monolinguals. Samuel et al., 2018 tested for the bilingual advantage in young adults and tested whether Korean culture could predict an advantage in the Simon task. In addition to the comparison of groups of bilinguals and monolinguals, they also compared performance in groups of Korean participants, British participants and a mixed cultural group. They found that Korean participants outperformed the British on global RT, global accuracy, and showed smaller Simon accuracy and RT effects. The study did not find any effect of bilingualism on performance on the Simon task, and when there was an effect, it was a bilingual disadvantage. These results mean that previous findings supporting a bilingual advantage with a greater proportion of East-Asian participants as bilinguals would need to be re-examined.

Measurement Problem in EF

Another set of shortcomings originate from the measurement of inhibition. Tasks commonly used to measure inhibition employ the calculation of difference scores. Difference scores are calculated as differences in reaction times for conditions where a conflict is present versus when it is absent. However, difference scores in tasks purported to measure the same inhibition construct do not significantly correlate (Kousaie and Phillips, 2012) or show only weak correlations. For instance,Paap and Greenberg, 2013 reported a correlation r = 0.01 between Simon and Flanker effects. Given mixed findings, and the lack of convergent and concurrent validity across tasks, it is important that when the 'bilingualism advantage in EF' hypothesis is examined, multiple measures of inhibition are used. Paap and Greenberg, 2013 recommend using atleast two measures. If interference scores in two tasks correlate, we can say that the two have a common mechanism for conflict resolution. (Paap, Anders-Jefferson, et al., 2020)

The Bayesian Way

Various researchers have called for the treatment of bilingualism as a set of continuous variables indexing the knowledge and use of multiple languages, and to incorporate the use of multiple measures of inhibition to provide a comprehensive view of the relationship between bilingualism and inhibition (Backer and Bortfeld, 2021). Studies responding to such calls have aimed at predicting inhibition performance from variables capturing bilinguals' language use and knowledge. This is done by building regression models with bilingualism-related characteristics as predictors of inhibition performance (operationalised as the difference in reaction times for incongruent and congruent trials) (Pot et al., 2018).

Since bilingualism involves many factors, a model containing all predictors is prone to overfit the data and consequently produce unreliable estimates of the regression coefficients. This limitation is typically resolved by reducing the set of predictors to a smaller subset of relevant factors. For instance, Pot et al., 2018 employed a partial least squares regression to find an "optimum subset" of predictor variables. Such an approach, however, ignores the uncertainty associated with the manner in which one arrives at the subset of relevant variables. This two-step process, where the first step involves reducing the number of predictor variables, and the second step consists in building a regression model with these selected predictors, can lead to misleading or biased inferences and overconfident parameter estimates. This is because a regression model is interpreted without evaluating the appropriateness of model selection techniques. (Bergh et al., 2021) With these caveats in mind, the current study examines the relationship between bilingualism and inhibition where we treat bilingualism as a set of continuous variables corresponding to language usage patterns and use Bayesian multi-model inference .

Bayesian Multi-model Inference Bayesian multi-model analyses allow us to carry out model selection and regression simultaneously. Here, we can calculate a "weight" for each model that captures how well the data supports a model. We simultaneously obtain predictions for each individual model. Then, predictions from all models are averaged by using model posterior probabilities as weights.By doing this, we can inspect the entire model space and the uncertainty present in each model. Variable selection and prediction occur together and allow us to overcome the limitations of a two-step process where the first step is illusorily treated as given *a priori*. (Bergh et al., 2021)

This process also overcomes the limitations of using R^2 as a yardstick for model comparison. Using the coefficient of determination to assess model fit is unsuitable for comparing models with different number of predictors. R^2 favours more complex models since R^2 can only increase as more predictors are added. Furthermore, complex models may overfit the data and treat noise as systematic relationships. In comparison, the Bayes Factor inherently penalizes complexity and is also a continuous measure of evidence.

Methods and Materials

Participants

Participants were selected if they met the following criteria: 1. they were English-Hindi bilinguals who were native-users of Hindi. 2. They were in the age group 16-35 years. 3. they have normal or corrected to normal vision.- Participants were recruited based on responses to calls for participation by

the institute via emails and posters. 66 Hindi-English bilinguals (*20 female and 46 male, mean age=22.7*) completed this experiment in two phases. One session included completion of the Language History Questionnaire 3.0 (Li et al., 2020) and the Bilingual Switching Questionnaire (Rodriguez-Fornells et al., 2012). This session had a duration of 30-45 minutes. The second session involved completion of four inhibition tasks. Participants completed the experiment online. Questionnaires were filled out online via google forms and the lhq-blclab website. Tasks were designed using JS-psych and were hosted on the institute server. Each block of the task took approximately 8-10 minutes. Participants were given a break of up to 10 minutes between tasks and within each task, they could take a break of 5 minutes between blocks. The protocol for this study was approved by the institute's IRB.

Materials

Measures of Bilingualism

- Language History Questionnaire 3.0: The LHQ 3.0 (Li et al., 2020) consists of 27 questions enquiring about demographic details and language use. The LHQ yields aggregate scores corresponding to proficiency in each language, dominance of each language (dominance ratios between languages), immersion in each language context and a multilingual diversity score (*MLDS*). The following aggregate scores were considered as independent variables: Proficiency L1, Proficiency L2, Immersion L1, Immersion L2, Dominance Ratio, and MLDS.

- Bilingual Language Switching Questionnaire (Rodriguez-Fornells et al., 2012): The BSWQ enquires about language switching frequency and includes 12 questions to tap four factors: (1) Switches into L1 (2) Switches into L2 (3) Contextual Switches (4) Unintended Switches. All responses were given on a 5-point Likert scale from "completely disagree" to "completely agree". A higher score indicates a higher switching frequency. An index of total switching frequency was calculated by summing reported switches into L1 and switches into L2. This index, along with frequencies of contextual and unintended switches were considered as independent variables.

Inhibition Tasks The experiment included four tasks that measure "inhibition" ability i.e., the ability to suppress task irrelevant information. The tasks differed on two domains- on the type of conflict involved (stimulus-stimulus vs. stimulus-response) and the nature of the stimulus used (verbal(words) vs. non-verbal(arrows)). In each task, a fixation (+) sign appeared for 500ms, followed by the stimulus (an arrow or word). The trial ended when the response was made. Participants were prompted to respond faster if response time exceeded 2 seconds. Each task consisted of two blocks. Within each block, 160 trials were divided into 120 congruent and 40 incongruent trials. Both blocks were preceded with 10 practice trials where feedback was provided. To prevent any ordering effects, the sequence of the four tasks was counterbalanced using a Latin square. The tasks were structured after Paap et al., 2019.

- Simon (Nonverbal): In the Simon task, arrows pointing up or down are presented on the screen. They appear either to the left or to the right of the fixation point. Subjects must ignore the location of the arrows, and instead respond (by pressing keys located on left and right of the keyboard) based on their orientation. Stimulus response conflict is seen in the Simon task as there is no overlap between the task relevant stimulus (direction in which the arrow is pointing) and task irrelevant stimulus (location of arrow on screen), since the arrows differ on a vertical plane, and the location differs on a horizontal plane. Participants respond to Arrow pointing UP by pressing the Z key and to Arrow pointing DOWN by pressing the M key.

- Simon (Verbal): In the verbal version of the Simon task, the arrows pointing up and down were replaced by the words "UP" and "DOWN" which appear on either side of the fixation. Participants respond to UP by pressing the Z key and to DOWN by pressing the M key.

- Vertical Stroop (Nonverbal): In the vertical Stroop task, stimulus-stimulus conflict is created by displacing the arrows presented on the vertical plane. Here, conflict is created between representations of location of arrow (above or below fixation) and type of arrow (pointing above or below). Participants respond to Arrow pointing UP by pressing the left arrow key and to Arrow pointing DOWN by pressing the right arrow key.

- Vertical Stroop (Verbal): In the verbal vertical Stroop task, the words "UP" or "DOWN" appear and are also displaced vertically. Participants respond to UP by pressing the right arrow key and to DOWN by pressing the left arrow key.

Data Preparation

For the four inhibition tasks, trials with RT greater than 3 standard deviations were removed before the calculation of mean reaction time for congruent and incongruent trials. For each participant and each task, Interference Effect scores were calculated as the difference between the mean RT on incongruent and congruent trials. The distribution of scores for the four tasks is presented in Figure 2.

Data Analyses

Based on our research questions and goals, we employed the following analyses: Multiple linear regression models were built for each task including bilingualism variables as predictors afnd interference scores as the dependent variable. Since the examination of the bilingual advantage warrants examination of the evidence for the null and the alternative, we undertook Bayesian multi-model analyses. Bayesian analysis can allow interpretations about the null, which null-hypothesis-testing regression does not allow. Four Bayesian multi-model

analyses were done. Bayesian regression was conducted using JASP. (JASP Team, 2022)

Estimation of regression coefficients Given a model M (a regression model with a subset of variables), regression coefficients are estimated. We start with prior beliefs about the values of regression coefficients represented by a probability distribution. For this study, we used the Jeffreys-Zellner-Siow (JZS) prior for the parameter β. This distribution is then updated using the Bayes theorem after observing data. The updated distribution rewards β values which predict data better by increasing the plausibility for those values. β values which predict data worse than average are penalized- their plausibility decreases.

Model Comparison The relative plausibility of each model is updated by using the data. Again, we start with some prior beliefs about the plausibility of each model. We employ the default model prior, which is a beta binomial prior with $\alpha = 1$ and $\beta = 1$. This prior ensures that models with equal number of predictors are equally likely. (Bergh et al., 2021) In our inferences, we consider what evidence exists for a specific model, which variables are important to predict inhibition performance and examine the regression coefficients for the predictors.

Table 1: Descriptive Statistics

Variable	Mean	S.D.	Min.	Max.
Proficiency $_{L1}$	0.825	0.107	0.57	1.00
Proficiency $_{L2}$	0.805	0.127	0.50	1.00
Immersion $_{L1}$	0.803	0.101	0.49	0.94
Immersion $_{L2}$	0.773	0.112	0.36	0.95
Dominance Ratio	1.031	0.292	0.34	2.51
MLDS	1.160	0.286	0.86	1.98
Unintended Switches	10.106	1.647	6.00	15.0
Contextual Switches	8.212	2.533	3.00	15.0
All Switches	17.182	3.369	8.00	25.0

Results

The descriptives for the bilingualism-related predictors employed in both frequentist and Bayesian regression are presented in Table 1. The average participant is a balanced bilingual with approximately equal levels of proficiency and immersion in both language contexts. We built four multiple regression models, one for each task. All four multiple linear regression models were non-significant (see Table 2).

Bayesian multiple regression To inform the current debate, we can compare the evidence supporting the null model (bilingualism variables do not predict performance), and the evidence supporting alternate models (all subsets with various combinations of the predictors) for each task. For the Simon tasks, the best model (model with the highest Bayes Factor for model odds) was the null model. Bayes Factor for model odds (BF_m) captures the change from the prior to posterior plausibility of models. For the non-verbal version of the Simon task, BF_m equalled to 13.113 and for the verbal version, BF_m equalled to 10.985. BF_{10} shows the relative predictive performance of the alternate model M_i and the null model M_0 for the obtained data. A BF_{10} value of 10 would indicate that the data are ten times more likely under M_i than M_0. For the non-verbal Simon task, BF_{10} for the alternate model containing all predictors was 0.008. Taking the reciprocal ($1/0.008 = 125$), we see that the observed data are 125 times more likely under the null. For the verbal Simon task, BF_{10} is 0.01, meaning that for this task, the data is 100 times more likely if the null were true rather than the alternate. Of course, it is possible that not all bilingualism variables are good predictors. The performance of the best 10 models which are made from subsets of all bilingualism predictors for the Simon tasks can be found in Table 4 and Table 5. All alternate models for the two tasks have BF_{10} values less than 3, comprising only anecdotal evidence. Similar findings are seen for the verbal vertical Stroop task, where the best model was the null model with $BF_m = 11.824$. Again, comparing the alternate model(which includes all predictors) and the null model, we see $BF_{10} = 0.01$. In contrast to the findings in these three tasks, for the non-verbal vertical Stroop task, the best model was the model containing only Multilingual Diversity Scores as a predictor ($BF_m = 15.629$). The data are 4 times more likely under this model than under the null ($BF_{10}) = 4.134$). This finding cannot be readily interpreted in favour of a "Bilingualism Effect", especially since these effects are not mimicked in the minimally different version of the same task. Intertask correlations are reported in Table 3.The intertask correlations demonstrate the lack of convergent validity of the Simon and Stroop tasks, prevalent in literature. For the minimally different versions of the tasks with the same kind of conflict, we see that there is only a weak correlation ($r = 0.214$) for the Simon and no significant correlation among the two Stroop tasks. We can also assess the explanatory ability of predictors by looking at the performance of models which include a specific predictor and models which do not include it. We can compare the prior probability (by summing prior probabilities of all models which include the variable) and the posterior probability (sum of probabilities of models which contain the variable after observing the data). The Bayes Factor for inclusion $BF_{inclusion}$ captures the change from the prior to posterior probability that a factor will be included in a model. We see that for all tasks, the probability of inclusion decreases for all factors after observing the data (except for multilingual diversity scores in the non-verbal Stroop analysis). The inclusion probabilities are presented in Figure 1.

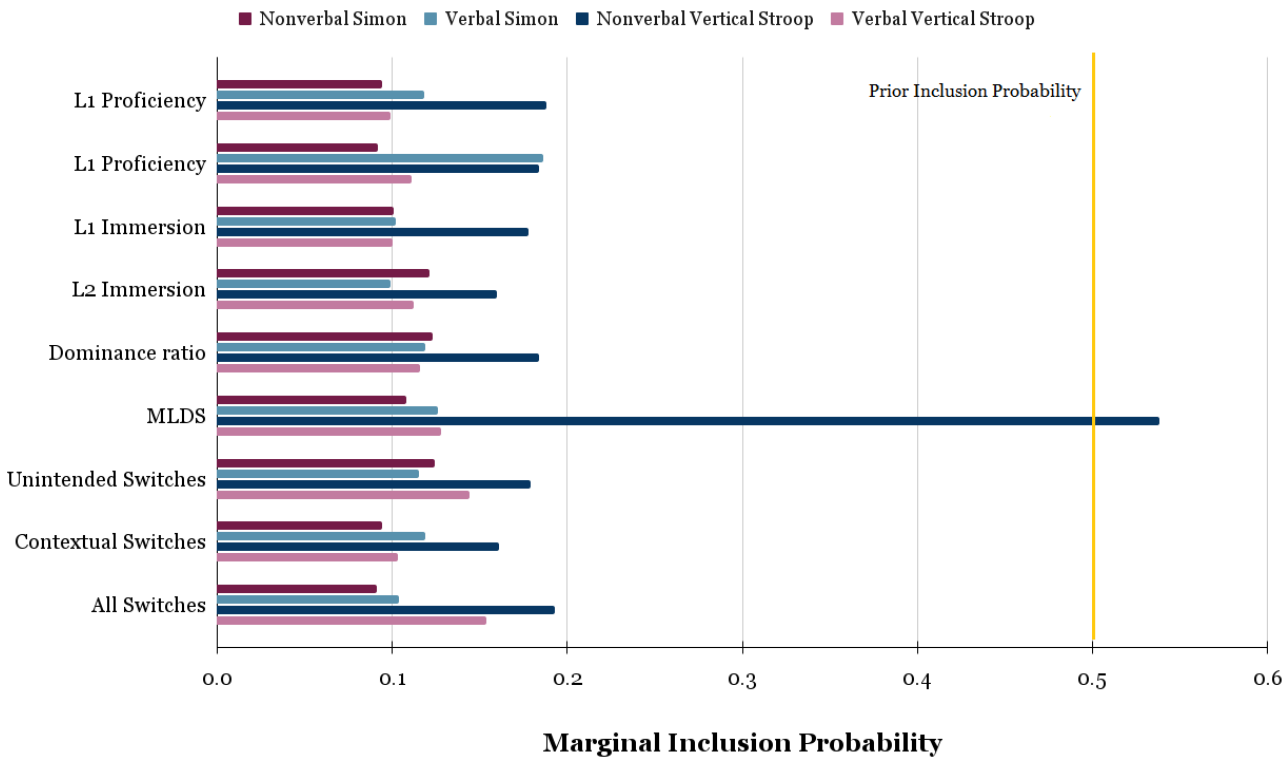

Figure 1: Updated Inclusion Probabilities. Prior inclusion probability = 0.5, represented by the yellow line.

Table 2: Linear Regression Models

Models	R	R^2	Adj. R^2	p
Simon (NV)	0.303	0.092	-0.054	0.769
Simon (V)	0.319	0.102	-0.043	0.701
Vertical Stroop (NV)	0.410	0.168	0.034	0.281
Vertical Stroop (V)	0.319	0.102	-0.043	0.702

Table 3: Pearson's Correlations for the four Inhibition Tasks

Variable	*NVSMN*	*VSMN*	*NVVS*	*VVS*
NVSMN	–			
VSMN	0.214∗	–		
NVVS	0.254∗	0.408∗∗	–	
VVS	−0.026	0.368∗∗	0.143	–

$^*\ p < 0.05$, $^{**}\ p < 0.001$

NVSMN:Nonverbal Simon, VSMN: Verbal Simon
NVVS: Nonverbal Vertical Stroop, VVS: Verbal Vertical Stroop

Discussion

In the current study, where we sought to evaluate the predictive efficacy of continuous measures of bilingualism for inhibition performance, we find that linear Regression models built for the four inhibition tasks failed to reach significance. At this juncture, an inference would be limited to say that we failed to find evidence that bilingualism affects inhibition. *"If the null is true, the best outcome of a significance test is a statement about a lack of evidence for an effect"* (Rouder et al., 2012). The debated question can be answered better by evaluating the evidence that exists *for* the null, which is made possible under the Bayesian framework. As described in the previous section, for three out of four tasks, the null model was the model which was best supported by the data. The dominant null findings are in line with a sizeable chunk of past null findings, such as that in Paap et al., 2019, who also report no significant relationships between bilingualism related variables (such as L2 Proficiency, Age of Acquisition, Language Switching Frequency) and inhibition. The claim that effects of bilingualism on inhibition would appear when we look at patterns of language knowledge and use (instead of group comparisons) , does not hold. We see that for the non-verbal vertical Stroop task, a "positive" finding appears. When examining the relationship between continuous bilingualism variables and inhibition, one should not make strong inferences that completely discard the null results observed for group comparisons. While the Multilingual Diversity score model has the greatest evidence for this task, it is not clear why. We cannot appeal to any theories which can

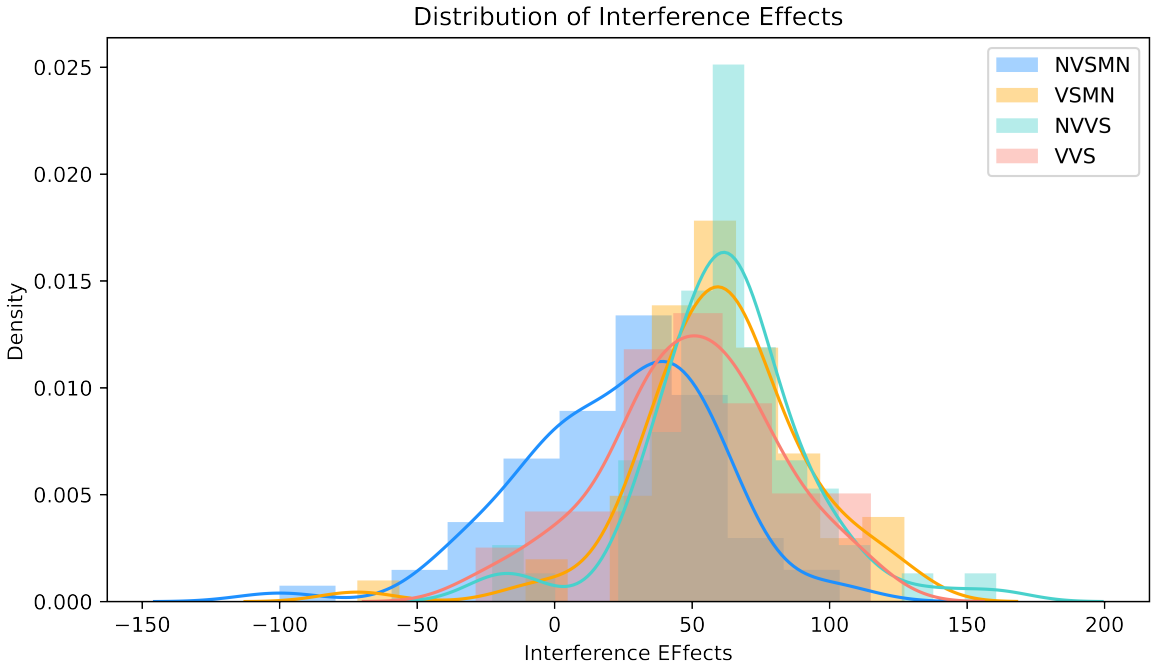

Figure 2: Distribution of Interference Effects for the four Inhibition tasks

describe or explain the golden conditions which existed for this task and made an effect of MLDS appear. We also cannot specify the necessary or sufficient conditions that would reliably reproduce this effect, since the task design was normative (structured after Paap et al., 2019) and no effects were seen on the verbal version of the same task. It is important to note that these findings occur in a broader context of null findings. Hence, the mixed results observed should not be interpreted in favor of "task-specificity of the bilingual advantage" but rather highlight the need to use multiple measures. When multiple measures of inhibition are used, null results dominate (Paap et al., 2016). Warned against "Type 1 incompetence" (lapses that cause disregard for null findings and confirmation bias driven false positives), in conjunction with the collected evidence, we are inclined to refute any claims of any facilitatory effects of bilingualism on inhibition.

Limitations

The conclusions made based on these results are marked by some limitations. One, measures related to bilingualism were based on self-reports only. In addition,the tasks we used purport to measure only one aspect of control - inhibition. More importantly, these tasks rely on difference-based scores and have questionable validity (Draheim et al., 2021; Paap, Anders-Jefferson, et al., 2020). These were used to allow comparisons to previous studies which employed them. Our conclusion that bilingualism has no facilitatory effects is also limited to the young adult population.

References

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L. J., Davidson, D. J., & Carreiras, M. (2014). Is there a bilingual advantage in the ant task? evidence from children. *Frontiers in psychology*, *5*, 398.

Backer, K. C., & Bortfeld, H. (2021). Characterizing bilingual effects on cognition: The search for meaningful individual differences. *Brain sciences*, *11*(1), 81.

Beatty-Martinez, A. L., Navarro-Torres, C. A., Dussias, P. E., Bajo, M. T., Guzzardo Tamargo, R. E., & Kroll, J. F. (2020). Interactional context mediates the consequences of bilingualism for language and cognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1022.

Bergh, D. v. d., Clyde, M. A., Gupta, A. R., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on bayesian multi-model linear regression with bas and jasp. *Behavior research methods*, 1–21.

Bialystok, E. (2011). Coordination of executive functions in monolingual and bilingual children. *Journal of experimental child psychology*, *110*(3), 461–468.

Bialystok, E., Barac, R., Blaye, A., & Poulin-Dubois, D. (2010). Word mapping and executive functioning in young monolingual and bilingual children. *Journal of cognition and development*, *11*(4), 485–508.

Blanco-Elorrieta, E., & Caramazza, A. (2021). On the need for theoretically guided approaches to possible bilingual advantages: An evaluation of the potential loci in the lan-

guage and executive control systems. *Neurobiology of Language*, *2*(4), 452–463.

Blumenfeld, H. K., & Marian, V. (2014). Cognitive control in bilinguals: Advantages in stimulus–stimulus inhibition. *Bilingualism: Language and Cognition*, *17*(3), 610–629.

Dash, T., & Kar, B. R. (2020). Behavioural and erp correlates of bilingual language control and general-purpose inhibitory control predicted by l1 and l2 proficiency. *Journal of Neurolinguistics*, *56*, 100914.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242.

JASP Team. (2022). JASP (Version 0.16.4)[Computer software].

Kousaie, S., & Phillips, N. A. (2012). Ageing and bilingualism: Absence of a "bilingual advantage" in stroop interference in a nonimmigrant sample. *Quarterly Journal of Experimental Psychology*, *65*(2), 356–369.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., De Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? a meta-analytic review. *Psychological bulletin*, *144*(4), 394.

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language history questionnaire (lhq3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938–944.

Paap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S., & Mason, L. (2019). On the encapsulation of bilingual language control. *Journal of Memory and Language*, *105*, 76–92.

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L., & Mikulinsky, R. (2020). Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, simon, and spatial stroop tasks? *Cognitive research: principles and implications*, *5*(1), 1–27.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive psychology*, *66*(2), 232–258.

Paap, K. R., Johnson, H. A., & Sawi, O. (2016). Should the search for bilingual advantages in executive functioning continue. *Cortex*, *74*(4), 305–314.

Paap, K. R., Mason, L., Zimiga, B., Ayala-Silva, Y., & Frost, M. (2020). The alchemy of confirmation bias transmutes expectations into bilingual advantages: A tale of two new meta-analyses. *Quarterly Journal of Experimental Psychology*, *73*(8), 1290–1299.

Pot, A., Keijzer, M., & De Bot, K. (2018). Intensity of multilingual language use predicts cognitive performance in some multilingual older adults. *Brain Sciences*, *8*(5), 92.

Rodriguez-Fornells, A., Krämer, U. M., Lorenzo-Seva, U., Festman, J., & Münte, T. F. (2012). Self-assessment of individual differences in language switching. *Frontiers in Psychology*, *2*, 388.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for anova designs. *Journal of mathematical psychology*, *56*(5), 356–374.

Samuel, S., Roehr-Brackin, K., Pak, H., & Kim, H. (2018). Cultural effects rather than a bilingual advantage in cognition: A review and an empirical study. *Cognitive Science*, *42*(7), 2313–2341.

Zirnstein, M., Bice, K., & Kroll, J. (2019). Variation in language experience shapes the consequences of bilingualism. *Bilingualism, Executive Function, and Beyond: Questions and Insights [Studies in Bilingualism, 57]*, 35–47.

Appendix A

Models	$P(M)$	$P(M\|data)$	BF_M	BF_{10}	R^2
Null model	0.100	0.593	13.113	1.000	0.000
DR	0.011	0.029	2.701	0.447	0.021
us	0.011	0.027	2.451	0.407	0.017
l2imm	0.011	0.024	2.227	0.370	0.014
mlds	0.011	0.023	2.118	0.353	0.012
l1prof	0.011	0.019	1.696	0.284	0.004
cs	0.011	0.018	1.589	0.266	0.002
l1imm	0.011	0.018	1.588	0.266	0.002
l2prof	0.011	0.017	1.528	0.256	0.001
alls	0.011	0.017	1.511	0.253	0.000
l1prof + l2prof + l1imm + l2imm + DR + mlds + us + cs + alls	0.100	0.005	0.043	0.008	0.092
l2imm + DR	0.003	0.004	1.450	0.244	0.037
DR + us	0.003	0.004	1.434	0.242	0.037
l1imm + l2imm	0.003	0.004	1.291	0.218	0.033
l2imm + us	0.003	0.003	1.228	0.207	0.031

Table 4: Best 10 models for the nonverbal Simon task. *l1prof: L1 Proficiency, l2prof: L2 Proficiency, l1imm : L1 Immersion, l2imm : L2 Immersion, DR: Dominance Ratio, us: Unintended Switches, cs: Contextual Switches, alls: All Switches, mlds: Multilingual Diversity Scores.*

Models	$P(M)$	$P(M\|data)$	BF_M	BF_{10}	R^2
Null model	0.100	0.550	10.985	1.000	0.000
l2prof	0.011	0.047	4.409	0.773	0.040
DR	0.011	0.024	2.218	0.398	0.017
cs	0.011	0.023	2.078	0.374	0.014
mlds	0.011	0.021	1.916	0.345	0.011
l1prof	0.011	0.019	1.679	0.303	0.007
us	0.011	0.018	1.614	0.292	0.005
l1imm	0.011	0.017	1.512	0.274	0.003
alls	0.011	0.016	1.472	0.266	0.002
l2imm	0.011	0.016	1.405	0.254	0.000
l2prof + mlds	0.003	0.006	2.344	0.425	0.057
l1prof + l2prof	0.003	0.006	2.299	0.417	0.056
l1prof + l2prof + l1imm + l2imm + DR + mlds + us + cs + alls	0.100	0.005	0.050	0.010	0.102
l2prof + us	0.003	0.005	1.773	0.322	0.047
l2prof + cs	0.003	0.005	1.691	0.307	0.045

Table 5: Best 10 models for the verbal Simon task. *l1prof: L1 Proficiency, l2prof: L2 Proficiency, l1imm : L1 Immersion, l2imm : L2 Immersion, DR: Dominance Ratio, us: Unintended Switches, cs: Contextual Switches, alls: All Switches, mlds: Multilingual Diversity Scores*