From oscillations to language: behavioural and electroencephalographic studies on cross-language interactions

Von Grebmer Zu Wolfsthurn, S.

CHAPTER 7

When left is right: The role of typological
similarity in multilinguals' inhibitory control
performance

Abstract: Both inhibitory control and typological similarity
between two languages feature frequently in current research on
multilingual cognitive processing mechanisms. Yet, the modulatory
effect of speaking two typologically highly similar languages on in-
hibitory control performance remains largely unexplored. However,
this is a critical issue because it speaks directly to the organisa-
tion of the multilingual's cognitive architecture. In this study, we
examined the influence of typological similarity on inhibitory con-
trol performance via a spatial Stroop paradigm in native Italian
and native Dutch late learners of Spanish. Contrary to our hypo-
thesis, we did not find evidence for a differential Stroop effect for
the typologically similar group (Italian-Spanish) compared to the

typologically dissimilar group (Dutch-Spanish). Our results therefore suggest a limited influence of typological similarity on inhibitory control performance. The study has critical implications for characterising inhibitory control processes in multilinguals.

Keywords: *inhibitory control performance, typological similarity, spatial Stroop task, late language learners*

7.1   Introduction

A remarkable feature of multilingual speakers is the ability to engage with several acquired languages, seemingly without effort. In this paper, we will broadly refer to multilinguals as those language users who have acquired one or more non-native language(s) in addition to their native language, L1 (Cenoz, 2013; De Groot, 2017). Over the past decades, numerous studies have attempted to capture the complexity of the multilingual experience. In particular, they focused on the cognitive, structural and functional consequences of managing several languages in the brain (Abutalebi & Green, 2007; Bialystok et al., 2012; D. W. Green, 1998; Kroll et al., 2015; Mosca & De Bot, 2017; Pliatsikas, 2020; Schwieter, 2016) (Sebastián-Gallés & Kroll, 2003).

A well-established aspect of the cognitive architecture of multilingualism is the parallel activation of languages across a range of proficiency levels, language combinations and linguistic domains (Blumenfeld & Marian, 2013; Colomé, 2001; Costa et al., 2000; Dijkstra, Van Heuven & Grainger, 1998; Guo & Peng, 2006; Hoshino & Thierry, 2011). In order to successfully mitigate parallel activation and to ultimately select the appropriate target language, multilinguals must employ a language control mechanism on the non-target language (Abutalebi & Green, 2007; Christoffels et al., 2007; Costa & Santesteban, 2004; Declerck et al., 2019; D. W. Green, 1998). Here, language control is conceptualised as a collection of control mechanisms applied to multilingual speech production and comprehension (Abutalebi, 2008; D. W. Green & Abutalebi, 2013). From

a theoretical point of view, this notion is featured in the Inhibitory
Control (IC) model of language control by D. W. Green (1998),
which postulates that the non-target language needs to be sup-
pressed prior to the linguistic output.

The exact nature of the mechanisms underlying language con-
trol is yet to be established. There is a substantial amount of evid-
ence suggesting that language control is strongly associated with
domain-general inhibitory control, also termed cognitive control or
executive control (Bialystok et al., 2012; Declerck et al., 2021; Fest-
man, Rodriguez-Fornells & Münte, 2010). Inhibitory control is an
executive function used to regulate and inhibit irrelevant inform-
ation with respect to thoughts or behaviour, as well as switching
attention (Diamond, 2013; Miyake et al., 2000). Some studies indic-
ate that language control impacts executive functions, for example
inhibitory control (Bialystok & Martin, 2004; Bialystok, 2010; Kroll
& Bialystok, 2013; D. W. Green & Abutalebi, 2013; Miyake et al.,
2000; Wiseheart, Viswanathan & Bialystok, 2016). Critically, evid-
ence further suggests that language control may share some un-
derlying processing mechanisms with inhibitory control (Declerck
et al., 2021; D. W. Green, 1998; Linck, Hoshino & Kroll, 2005;
Weissberger, Gollan, Bondi, Clark & Wierenga, 2015), although
this notion is still debated (Branzi, Della Rosa, Canini, Costa &
Abutalebi, 2016; Calabria, Hernandez, Branzi & Costa, 2012).

In the current study, we investigated the impact of multilin-
gualism on inhibitory control performance. More specifically, we
examined whether the typological similarity between languages of
a multilingual plays a role in modulating inhibitory control per-
formance. Typological similarity, also termed typological distance
or language similarity in the literature, refers to linguistic and struc-
tural (dis)similarities across different languages spoken by multilin-
guals (Foote, 2009; Putnam et al., 2018; Westergaard, Mitrofanova,
Mykhaylyk & Rodina, 2017). For example, Italian and Spanish may
be considered as more typologically similar languages compared to
language pairs such as Dutch and Spanish because of the larger
degree of overlap in morphosyntax, gender systems and cognates

(Paolieri et al., 2019; Schepens et al., 2012; Serratrice et al., 2012).

Several studies have focused on the modulating effects of typological similarity on language control, for example, within the context of a classical Stroop paradigm (Brauer, 1998; Coderre, Van Heuven & Conklin, 2013; Van Heuven, Conklin, Coderre, Guo & Dijkstra, 2011) However, studies directly investigating the effect of typological similarity on domain-general inhibitory control performance are scarce, but see Bialystok et al. (2005), Linck et al. (2005) and Yamasaki et al. (2018). Typical experimental paradigms to explore domain-general inhibitory control are the Simon task (Bialystok, Craik, Klein & Viswanathan, 2004; Bialystok et al., 2005; Simon & Small, 1969), and the spatial Stroop task (Hilbert et al., 2014; Lu & Proctor, 1995; Luo & Proctor, 2013). The core feature of the Simon task is a conflict between the physical location of a stimulus and the response, e.g., a stimulus appearing on the right side of a screen while the corresponding response button is located on the left side. The Simon effect quantifies the difference in response times (RTs) between trials in which stimulus and response location match and trials in which stimulus and response location mismatch. Typically, longer RTs are linked to the mismatch trials. Accordingly, a smaller Simon effect reflects better inhibitory control performance, whereas a larger Simon effect reflects lower inhibitory control performance (Bialystok et al., 2004).

In this study, we used the spatial Stroop task (Hilbert et al., 2014; Lu & Proctor, 1995), which is a combination of the Simon task and the classical colour-word Stroop task (MacLeod, 1992; Stroop, 1935). While the classical Stroop task involves the naming of a colour-word written in either the matching ink colour (congruent trial), e.g., the word *RED* written in red ink, or the mismatching ink colour (incongruent trial), e.g., the word *RED* written in blue ink, the spatial Stroop task focuses on spatial stimulus-stimulus conflicts. The basic feature of the spatial Stroop task is that a target word ("left", "right", "up", "down") either matches its location on the screen, e.g., *LEFT* shown on the left side of the screen (*congruent* trial), or it does not match its location on the screen, e.g.,

LEFT shown on the right side of the screen (*incongruent* trial). The key to success in this task is to inhibit the irrelevant spatial stimulus property (e.g., the location of the word) and to instead focus on the relevant target stimulus property (the target word itself). In this task, inhibitory control performance is reflected in the spatial *Stroop effect*, which describes the quantitative difference in RTs between *congruent* and *incongruent* trials (Hilbert et al., 2014; La Heij, Van der Heijden & Plooij, 2001; Marian et al., 2013; Roelofs, 2021; Van Heuven et al., 2011). Drawing parallels between the Simon task, a smaller Stroop effect is reported to indicate better inhibitory control performance (Costa, Hernández & Sebastián-Gallés, 2008; Heidlmayr et al., 2014; Pardo, Pardo, Janer & Raichle, 1990).

In the current study, the critical question we sought to answer was the following: does typological similarity between the two languages significantly modulate inhibitory control performance in multilinguals? A relevant theoretical framework for this particular question is the Conditional Routing Model (CRM) by Stocco et al. (2014). The model is based upon the notion that the multilingual experience dynamically impacts domain-general executive functions, including inhibitory control, as a result of the parallel activation of the languages (Bialystok & Martin, 2004; Festman et al., 2010). Here, the model postulates that executive functions are effectively trained over time (Kroll & Bialystok, 2013; Yamasaki et al., 2018), which results in a strengthening of the neural circuits underlying these executive functions. When the languages within a multilingual system are highly typologically similar, one may predict a higher degree of cross-language interference (Cenoz, 2001; J. Chen, Zhao, Zhaxi & Liu, 2020; De Bot, 2004). In turn, this implies that speakers of these languages develop better inhibitory control skills compared to speakers of typologically less similar languages (Yamasaki et al., 2018). Therefore, the CRM provides us with a testable prediction for the effect of typological similarity on inhibitory control performance: speakers of typologically similar languages should exhibit a better inhibitory control performance compared to speakers of typologically less similar languages. Ap-

plied to the context of a spatial Stroop task used in this study, speakers of typologically similar languages (e.g., Italian-Spanish) should therefore show a smaller Stroop effect compared to speakers of typologically less similar languages (e.g., Dutch-Spanish speakers).

7.1.1   The current study

We explored the modulatory role of typological similarity on inhibitory control performance in a spatial Stroop task (hereafter simply *Stroop task*) in two groups of speakers with differing degrees of typological similarity. Participants were native Italian learners of Spanish, and native Dutch learners of Spanish. On the basis of typological work by Schepens et al. (2012) and Van der Slik (2010), we defined our Italian-Spanish group as our typologically similar group, and our Dutch-Spanish group as our typologically dissimilar group. All participants had a Spanish proficiency level in the B1/B2 range within the CEFR framework (Council of Europe, 2001). We followed a spatial Stroop paradigm inspired by Hilbert et al. (2014), who used the location words "left", "right", "up" and "down" to study the Stroop effect in native speakers of German: see also Lu and Proctor (1995) and Shor (1970). In our paradigm, we exploited the conflict between the target word and the location of the target word on the screen, for example, the Spanish location word [*izquierda*] "left" displayed on the right side of the screen, or the Spanish word [*derecha*] "right" displayed on the left side of the screen. The translation equivalents for [*izquierda*] "left" and [*derecha*] "right" are "sinistra" and "destra" in Italian, and "links" and "rechts" in Dutch, respectively. In the *congruent* condition, the target word and the target word location matched. In contrast, in the *incongruent* condition, the target word and the target word location did not match. We measured accuracy and RTs during this task. Post-experiment, we calculated the Stroop effect by subtracting the RTs for congruent trials from RTs for incongruent trials. Importantly, we employed an equiprobable Stroop task design, whereby the probability of each condition occurring in the subsequent trial is identical. Within the framework of the Dual Mechanisms of Control

(DMC) model (Braver, 2012), an equiprobable Stroop task design is linked to a *proactive* control strategy. At the core of this particular strategy is the maintenance of goal-relevant information over time to succeed at the task (Braver, 2012; Gonthier, Braver & Bugg, 2016). Therefore, our Stroop task taps not only into inhibitory control performance per se, but also into the cognitive mechanisms of monitoring the task.

Research questions

Our research questions were the following: first, is there a difference in terms of RTs as a function of typological similarity (typologically similar vs. typologically dissimilar)? Secondly, connected to this first question, is the Stroop effect larger for one group compared to the other, thereby reflecting an effect of typological similarity on inhibitory control performance?

Hypotheses

Based on the literature outlined above, we first predicted a Stroop effect for both the Italian-Spanish group and the Dutch-Spanish group. Behaviourally speaking, this would be reflected in higher accuracy and shorter RTs for congruent trials compared to incongruent trials. Next, in line with the CRM (Stocco et al., 2014), we hypothesised overall shorter RTs for the Italian-Spanish group compared to the Dutch-Spanish group. Finally, we expected a difference in inhibitory control performance as a function of typological similarity: we expected an interaction effect of condition (congruent vs. incongruent) and typological similarity (typologically similar vs. typologically dissimilar) on the size of the Stroop effect. A smaller Stroop effect for the Italian-Spanish group would imply that the overall inhibitory control performance is better for the typologically similar languages compared to the less typologically similar Dutch-Spanish group. In turn, this would support the CRM (Stocco et al., 2014).

7.2   Methods

In addition to the spatial Stroop task, we asked participants to complete the Language Experience and Proficiency Questionnaire, LEAP-Q (Marian et al., 2007). The LEAP-Q is a questionnaire designed to obtain a measure for the linguistic profile of our participants in terms of their proficiency levels and experiences with the languages within their multilingual system (Marian et al., 2007). Finally, participants also completed the LexTALE-Esp (Izura et al., 2014), a lexical decision task to establish vocabulary size in Spanish, for descriptive purposes.

7.2.1   Participants

For the Italian-Spanish group, we recruited 33 healthy, right-handed native speakers of Italian (24 females) with a B1/B2 level of Spanish at Pompeu Fabra University (Barcelona, Spain). Mean age of the Italian-Spanish group was 27.12 years ($SD = 4.08$). Our recruitment criteria for this group were the following: no additional language learnt before the age of three, age of acquisition of Spanish from fourteen years onwards, a maximum time spent in a Spanish-speaking country of no longer than one year, no psychological, neurological, visual, auditory, or language-related impairments; and finally, an age range between 18 and 35 years. For the Dutch-Spanish group, we recruited and tested 25 healthy, right-handed native speakers of Dutch (16 females) with a B1/B2 level of Spanish at Leiden University (Leiden, The Netherlands). Mean age of the Dutch-Spanish group was 22.84 years ($SD = 3.05$). Our recruitment criteria were identical to the Italian-Spanish group, with the cap on maximum time spent in a Spanish-speaking country less stringent due to the testing location. Data from the LEAP-Q was analysed to establish a detailed linguistic profile of each participant. See Appendix 7.A and Appendix 7.B for an overview of the profiles for the Italian-Spanish group and the Dutch-Spanish group, respectively.

LEAP-Q: Linguistic profile of participants

Italian-Spanish group. With respect to their linguistic profile in Spanish, two participants acquired Spanish as first foreign language, whereas eighteen participants acquired Spanish as second foreign language. Spanish was the third foreign language for ten participants, and three participants acquired Spanish as fourth foreign language (Appendix 7.A). The mean age of acquisition (AoA) of Spanish was 23.93 years ($SD = 5.07$). On average, participants reported to be fluent in Spanish at the age of 24.88 years ($SD = 4.48$), to have started reading in Spanish at the age of 24.36 years ($SD = 4.91$) and to be fluent readers by the age of 24.24 ($SD = 4.82$). On average, participants spent 0.46 years ($SD = 0.343$) in a Spanish-speaking country and had learnt Spanish for 0.93 years ($SD = 1.17$) either at school as a foreign language, or as a language course in Spain. Twenty-five participants were completing or had completed a formal Spanish language course that was not part of the school curriculum shortly before or upon their arrival in Spain (mean length of course: 0.53 years, $SD = 0.889$ years). Finally, participants quantified their current daily exposure to Spanish as 40% ($SD = 18.37\%$) of the time with respect to the other languages spoken. In terms of dominance, thirteen participants classified Spanish as their most dominant language after Italian, fourteen participants as their second most dominant language after Italian, five participants as their third most dominant language after Italian, and one participant as their fourth most dominant language after Italian. On a ten-point scale, ten being maximally proficient, participants rated their speaking proficiency at 6.09 ($SD = 1.76$), comprehension proficiency at 7.26 ($SD = 1.67$) and their reading proficiency at 7.36 ($SD = 1.48$).

Dutch-Spanish group. In this group, nine participants stated that they acquired Spanish as their second foreign language, nine participants as their third foreign language and seven participants as their fourth foreign language (Appendix 7.B). Mean AoA of Spanish was 17.84 years ($SD = 3.16$). People stated to be fluent in Spanish on average at the age of 19.60 years ($SD = 2.52$), that

they started reading in Spanish at the age of 18.44 years ($SD =$ 3.24), and that they were on average fluent in reading by 19.76 years ($SD = 3.41$). Eighteen out of the twenty-five participants spent on average 0.57 years ($SD = 0.66$) in a Spanish-speaking country (e.g., Spain, Argentina, Colombia, Mexico). Compared to the other languages, participants quantified their daily exposure to Spanish with 12.96% ($SD = 10.07$). Critically, two participants reported Spanish as their second most dominant language, nineteen as their third most dominant, three as their fourth most dominant and one participant as their fifth most dominant language following Dutch. On a ten-point scale (ten being maximally proficient), participants reported an average speaking proficiency in Spanish of 6.40 ($SD = 1.47$), a comprehension proficiency of 7.08 ($SD = 1.32$) and a reading proficiency of 7.08 ($SD = 1.22$). These ratings are highly comparable with the Italian-Spanish group.

7.2.2   Materials and design

Prior to the experiment, participants completed the LEAP-Q (Marian et al., 2007) at home to reduce self-report biases frequently induced in laboratory settings (Rosenman et al., 2011). During the experiment, we first asked participants to complete the LexTALE-Esp (Izura et al., 2014), followed by the Stroop task.

Tasks and stimuli

LexTALE-Esp. We administered the LexTALE-Esp to establish vocabulary size in Spanish. The task was programmed in E-prime2 (Schneider et al., 2002), using the exact same stimuli as in the original version by Izura et al. (2014).

Stroop task. We administered the Stroop task to measure inhibitory control performance in our Italian-Spanish speakers and Dutch-Spanish speakers. We again generated an E-prime2 script (Schneider et al., 2002) for this task. The target words were the written Spanish words [*izquierda*] "left" and [*derecha*] "right".

7.2.3   Procedure

Prior to initiating the experiment, participants were provided with an information sheet and the opportunity to ask clarification questions. Then, participants signed the consent form in compliance with the ethics code for linguistic research at the Faculty of Humanities at Leiden University. Before each task, we provided participants with written task instructions in Spanish. Upon termination of all tasks, participants were provided with a debrief sheet in their respective L1, they signed the final consent form and received a monetary compensation for their participation.

LexTALE-Esp

The LexTALE-Esp procedure was identical for both groups. We asked participants to indicate via a button press whether the string corresponded to a Spanish *word* (e.g., [*secuestro*] "kidnapping") or a *pseudoword* (e.g., *plaudir*). Participants were instructed that incorrectly assigning a word status to a pseudoword and vice versa would lead to a deduction in the score. The trial procedure was as follows: first, a black fixation cross was displayed for 1,000 ms on a white screen. Then, a letter string corresponding to either a word or a pseudoword was displayed in the centre of the screen. The letter string remained on the screen until the participants' response. After the participants' response, the next trial was initiated. Sixty trials were Spanish word trials, whereas thirty were pseudoword trials. Trial order was randomised so that each participant was presented with a unique trial order.

Stroop task

The procedure for the Stroop task was the same for both groups. Participants were asked to focus on the target word while ignoring the location of the target word on the screen and to respond to the target word via button presses. The trial procedure was as follows: first, participants saw a black fixation cross for 500 ms in the centre of a white screen. Next, they saw a target word appear on either the left or right side of the screen along the horizontal midline in

Spanish. This target word was either [*izquierda*] "left" or [*derecha*] "right". The target word was visible on the screen until participants responded or for a maximum display time of 1,000 ms (Figure 7.2.1).

Figure 7.2.1: *Example trial procedure for a congruent trial followed by an incongruent trial.*
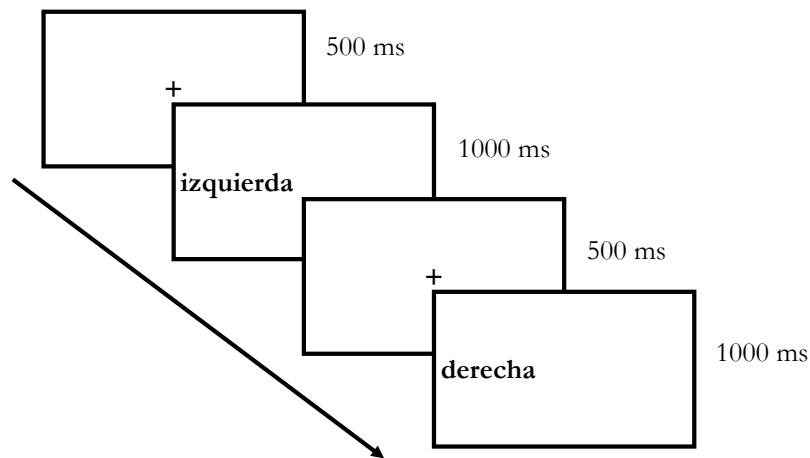
The next trial was initiated after participants' response, or if the response time limit was reached. Half of the trials were *congruent* trials, where the target word matched the location on the screen. The other half of the trials were *incongruent* trials, where the target word and the location on the screen did not match. There were 24 trials for each target word (*izquierda/ derecha*) and target location on the screen (left side/right side), amounting to 48 trials for the congruent condition and 48 trials for the incongruent condition. Prior to the start of the main experimental round, there was a short practise round to familiarise participants with the task procedure. Trial order was randomized in the practice round and in the main experimental round.

7.3 Results

7.3.1 Data exclusion

For the Italian-Spanish group, data from one participant were lost due to a technical failure. Therefore, we included 32 datasets in the analysis. In contrast, for the Dutch-Spanish group we included all 25 datasets in the analysis, adding to a total of 57 datasets.

7.3.2 Data analysis

We analysed our behavioural data using R, Version 4.0.3 in RStudio, Version 1.4.1106 (R Core Team, 2020). We employed a single trial generalised linear mixed effects modelling approach using the *lme4* package (Bates et al., 2020). We first modelled the outcome variables *accuracy* and *RTs* separately for each individual group. Next, we pooled our RT data from both groups for a group comparison analysis to study potential effects of typological similarity on *Stroop effect sizes*. For both the individual group analyses and the group comparison analysis, we applied the following model fitting procedure: first, we constructed a theoretically plausible model with a maximal random effects structure as supported by our data (Barr, 2013; Matuschek et al., 2017). In our case, the maximal model was a random-intercept and random-slope model for both accuracy and RTs. In the case of non-convergence or singular fit, we simplified our random effects structure. Next, we generated the model of best fit in a top-down procedure, whereby we simplified the fixed effects structure in a stepwise fashion. After fitting each model, we performed model diagnostics to establish the goodness of fit using the *DHARMa* package (Hartig, 2020). This involved the plotting of the model residuals against the predicted values, and closely investigating the distribution of the residuals and the presence of influential data points to identify issues in terms of the model fit. Then, we compared models with different fixed effects structures to establish the model of best fit using the *anova()* function, which is based on the Akaike's Information Criterion, AIC (Akaike, 1974), the Bayesian Information Criterion,

BIC (Neath & Cavanaugh, 2012) and the log-likelihood ratio. To test for the significance of the terms in the fixed effects structure, absolute test-statistics greater than 1.96 were interpreted as statistically significant at $\alpha = 0.05$ (Alday et al., 2017; Matuschek et al., 2017). Finally, the models of best fit for RTs were re-fitted using the REML criterion (Bates, Mächler, Bolker & Walker, 2014; Verbyla, 2019). All best-fitting models and model parameters are reported in Appendix 7.C, Appendix 7.D and Appendix 7.E. Note that the model parameters for accuracy are reported as odds ratios.

To model accuracy, we used the *glmer()* function with a binomial distribution. This particular function from the *lme4* package uses maximum likelihood estimation via the Laplace approximation (Bates et al., 2020). In contrast, we used the *lmer()* function with a normal distribution to model RTs for correct trials. For the individual group analysis, our fixed effect of interest was *condition* (congruent vs. incongruent), whereas *subject* and *item* were included as random effects. For the group comparison analysis, we used the *lmer()* function to model the interaction effect of *condition* (congruent vs. incongruent) and *typological similarity* (typologically similar vs. typologically dissimilar) as well as their main effects on RTs. *Subject* and *item* were again included as random effects. To control for potential covariates, we included *LexTALE-Esp score* and *order of acquisition of Spanish* as fixed effects in all analyses.

7.3.3   LexTALE-Esp

Post-experiment, we calculated the LexTALE-Esp vocabulary size score for each participant by subtracting the percentage of incorrectly identified pseudowords from the percentage of correctly identified words (Izura et al., 2014). For the Italian-Spanish group, the mean LexTALE-Esp score was 26.30 ($SD = 14.04$). Large individual differences were evident from the range of scores, which was between -7.37 to 49.30. In contrast, the mean LexTALE-Esp score for the Dutch-Spanish group was 22.69 ($SD = 17.19$). The range was from -11.92 to 54.73, which yielded similar large individual dif-

ferences between participants. A two-sample t-test yielded no significant statistical difference in LexTALE-Esp scores between the two groups with $t(45.90) = 0.851$, $p = 0.399$. According to calculations provided by Lemhöfer and Broersma (2012), all of our speakers were at or below the B2 level for Spanish according to CEFR standards (Council of Europe, 2001), in line with our recruitment criteria.

7.3.4   Stroop task

We first computed descriptive statistics for accuracy and RTs for both groups. See Table 7.3.1 for descriptive mean accuracy, mean RTs and Stroop effects for the Italian-Spanish group and the Dutch-Spanish group. Descriptively speaking, results yielded overall longer RTs for the Italian-Spanish group compared to the Dutch-Spanish group. Moreover, the Stroop effect was descriptively larger for the typologically similar languages compared to the typologically dissimilar languages. We first discuss the individual analysis for the Italian-Spanish group and the Dutch-Spanish group, respectively. Then, we discuss the group comparison for the Stroop effect.

Table 7.3.1: *Mean accuracy and RTs for the Stroop task for the Italian-Spanish group (n = 32) and the Dutch-Spanish group (n = 25).*

	Italian		Dutch	
	Accuracy (%)	**RTs (ms)**	**Accuracy (%)**	**RTs (ms)**
Condition	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
congruent	94.73 (22.35)	579 (113)	95.67 (20.37)	560 (123)
incongruent	91.21 (28.32)	607 (112)	90.67 (29.10)	576 (112)
Stroop effect		**28**		**16**

Italian-Spanish group

Accuracy. For the Italian-Spanish group, the model of best fit included *condition* as fixed effect, as well as *subject* and *item* as random effects. The by-*subject* random slopes for *condition* led to singular fit and were therefore dropped from the model fitting procedure. The covariates *LexTALE-Esp score* and *order of acquisition of Spanish* did not significantly improve the model fit. Therefore, the model of best fit was the following: accuracy $\sim$ condition (congruent vs. incongruent) + (1|subject) + (1|item). Participants were significantly more accurate for congruent trials compared to incongruent trials with $\beta = 0.560$, *95% CI*[0.400, 0.784], $z = -3.38$, $p = 0.001$ (see Appendix 7.C for the full model parameters). See Figure 7.3.1 for mean accuracy for the Italian-Spanish group.

Response times. For the Italian-Spanish group, the model of best fit yielded an effect of *condition*, a random effect for *subject* and *item* and a by-*subject* random slope for *condition* (Figure 7.3.1). Neither *LexTALE-Esp score* nor *order of acquisition of Spanish* significantly modulated the outcome variable or improved the model fit. These two covariates were therefore excluded from the model fitting procedure. This resulted in the following best-fitting model: RTs $\sim$ condition (congruent vs. incongruent) + (condition|subject) + (1|item). Participants were statistically faster in responding in the congruent condition compared to the incongruent condition with $\beta = 29.46$, *95% CI*[18.10, 40.82], $t = 5.09$, $p < 0.001$ (see Appendix 7.C).

Figure 7.3.1: *Mean accuracy (A) for each participant and response times (B) for each condition for the Italian-Spanish group (n = 32).*
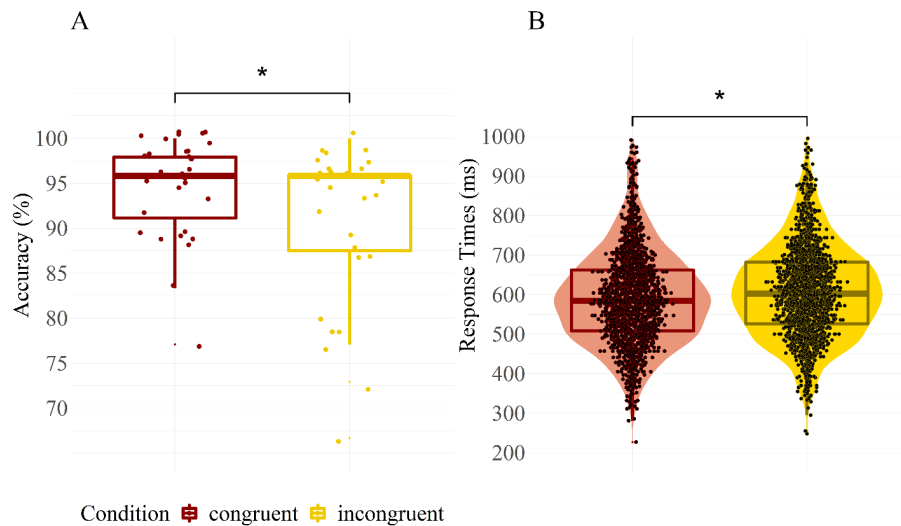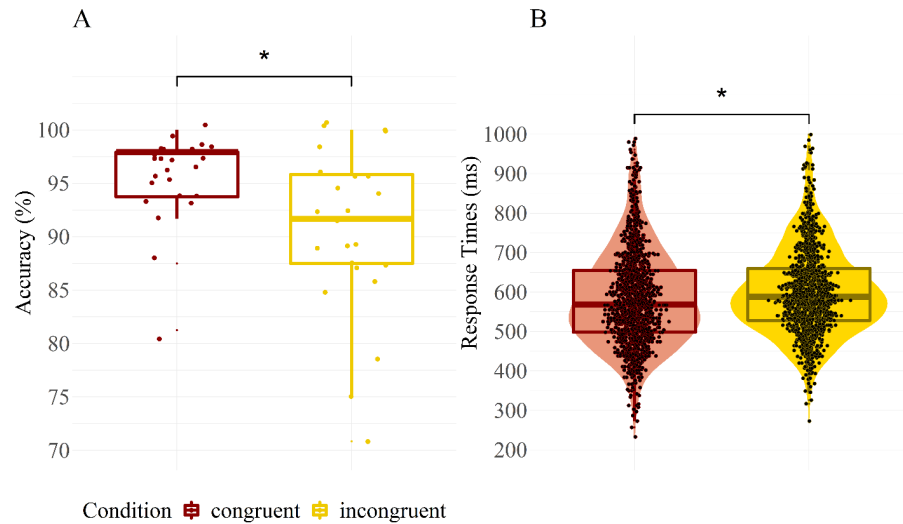

Dutch-Spanish group

Accuracy. For the Dutch-Spanish group, the model of best fit included a fixed effect of *condition*, as well as by-*subject* random slopes for *condition* and *subject* as random effect. *Item* led to singular fit and was excluded from the model fitting procedure. Further, *Lextale-Esp score* and *order of acquisition of Spanish* did not significantly improve the model fit. The model of best fit was as follows: accuracy ~ condition (congruent vs. incongruent) + (condition|subject). Participants were significantly more accurate in the congruent compared to the incongruent condition with β = 0.477, *95% CI*[0.295 – 0.771], z = -3.02, p = 0.003 (see Appendix 7.D for the full model parameters). See Figure 7.3.2 for mean accuracy for the Dutch-Spanish group.

Response times. For the Dutch-Spanish group, we found that the model of best fit included *condition* as fixed effect, *subject* as random effect and by-*subject* random slopes for *condition* (Figure 7.3.2). The random effect for *item* was not supported by our data and was therefore excluded from the random effects structure.

Neither *LexTALE-Esp score* nor *order of acquisition of Spanish* significantly improved the model fit and were subsequently dropped from the model selection procedure. The resulting model of best fit was the following: RTs $\sim$ condition (congruent vs. incongruent) + (condition|subject). Participants were significantly faster in responding in the congruent condition compared to the incongruent condition, with $\beta = 15.31$, *95% CI*[4.40, 26.21], $t = 2.75$, $p = 0.006$ (see Appendix 7.D).

Figure 7.3.2: *Mean accuracy (A) for each participant and response times (B) for each condition for the Dutch-Spanish group (n = 25).*

Taken together, data from both the Italian-Spanish group and the Dutch-Spanish groups suggest that participants were significantly more accurate and faster in the *congruent* condition compared to the *incongruent* condition. Therefore, both groups displayed the Stroop effect.

Stroop effect: group comparison

Finally, we compared the Stroop effect (RT incongruent trials minus RTs congruent trials) between the Italian-Spanish and Dutch-Spanish group to explore the possible impact of typological

similarity. Here, we explored the interaction effect between *condition* and *typological similarity* on the size of the Stroop effect. Descriptively speaking, the Stroop effect was larger for the Italian-Spanish group compared to the Dutch-Spanish group. However, the model of best fit yielded a main effect of *condition* with participants being faster for *congruent* trials compared to *incongruent* trials with $\beta = 23.20$, *95% CI*[14.56, 31.83] = 4.40, $t = 5.27$, $p < 0.001$. The model also included a main effect of *typological similarity*, with participants from the typologically similar group (Italian-Spanish) being significantly slower compared to the typologically dissimilar group (Dutch-Spanish) with $\beta = 30.70$, *95% CI*[7.72, 53.68] = 11.72, $t = 2.62$, $p = 0.009$. Therfore, the best-fitting model was: RTs ~ typological similarity (high vs. low) + condition (congruent vs. incongruent) + (condition|subject) + (1|item), see Appendix 7.E. There was no evidence for an interaction effect between *condition* and *typological similarity*. See Appendix 7.E for full model specification details, as well as a comparison between the model that included the interaction term and the best-fitting model that did not include the interaction term. Further, the model of best fit also included a by-*subject* random slope for *condition* as well as *item* as random effect. The covariates *LexTALE-Esp score* and *order of acquisition of Spanish* did not significantly contribute to improving the model fit and were therefore not included in the final model. See Figure 7.3.3 for the comparison of the Stroop effect across the Italian-Spanish and Dutch-Spanish group.

Figure 7.3.3: *Mean response times for the Italian-Spanish group (left) and the Dutch-Spanish group (right) for each condition for Spanish Stroop targets (n = 57).*
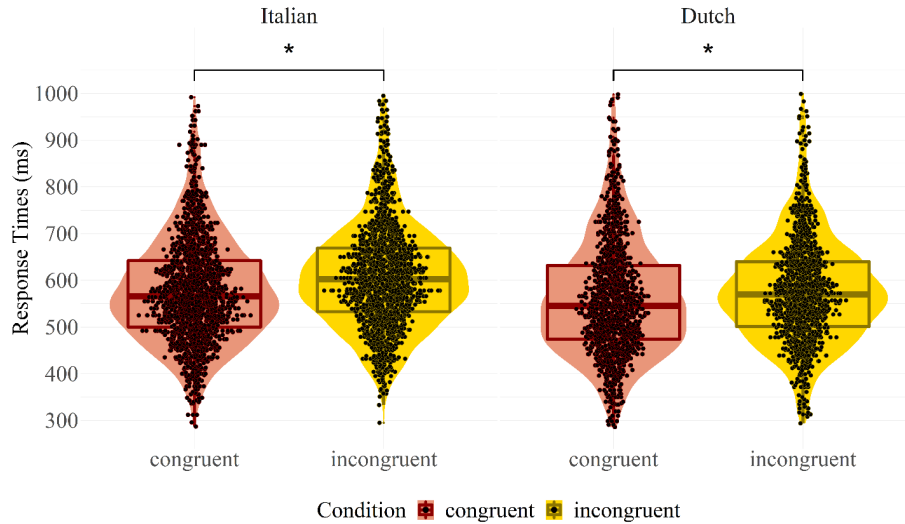
7.4   Discussion

In this study, we explored the effect of typological similarity on inhibitory control performance in a group of Italian-Spanish speakers and a group of Dutch-Spanish speakers via a spatial Stroop task. The goal of this study was twofold: first, we examined whether or not the typologically similar (Italian-Spanish) group showed a general processing advantage over the typologically dissimilar (Dutch-Spanish) group in terms of RTs. Secondly, we studied whether typological similarity yielded a difference between the two groups in terms of the Stroop effect (difference in RTs between incongruent and congruent trials). Here, a smaller Stroop effect would be indicative of better inhibitory control performance. On the basis of the CRM (Stocco et al., 2014), we expected shorter RTs and a smaller Stroop effect for the Italian-Spanish group compared to the Dutch-Spanish group.

Stroop data from the Italian-Spanish as well as the Dutch-

Spanish group showed that participants were sensitive to the inherent task conflict. More specifically, results demonstrated higher accuracy and shorter RTs for congruent compared to incongruent trials. This yields the typical *Stroop effect*, which is a measure of inhibitory control performance in this task. To succeed at this task, participants had to ignore the irrelevant information (i.e., the location of the target) and instead focus on the target word itself to provide a correct response. Further, as discussed in the introduction, participants had to employ a proactive control strategy (Braver, 2012; Gonthier et al., 2016) and monitor the goal-relevant information during the task, as described in the DMC model (Braver, 2012). Therefore, the presence of a Stroop effect in both groups reflects not only a measure for inhibitory control performance, but also a monitoring strategy to solve this task.

With respect to the first research question, the group comparison analysis showed that the typologically dissimilar (Dutch-Spanish) group was comparatively faster than the typologically similar (Italian-Spanish) group in this task. This finding contrasts with our predictions. The original prediction on the basis of the CRM (Stocco et al., 2014) was a processing advantage for the typologically similar Italian-Spanish group compared to the Dutch-Spanish group due to continuous training of executive functions and inhibitory control skills over time. In contrast, our findings suggest that typologically dissimilar Dutch-Spanish group had a processing advantage in terms of RTs over the Italian-Spanish group. In the literature, similar findings were reported by Bialystok et al. (2005), who investigated the role of typological similarity on the performance during a Simon task in highly proficient Cantonese-English speakers (typologically dissimilar group) and highly proficient French-English speakers (typologically more similar group). Results showed a processing advantage for Cantonese-English speakers compared to the French-English speakers in the form of faster RTs on the Simon task for Cantonese-English speakers, see also Linck et al. (2005). Our results are comparable to Bialystok et al. (2005), and suggest that in this particular task, typological dissimilarity was advantageous over typological similarity. Moreover, these results suggest

a qualitative difference between the Italian-Spanish and the Dutch-Spanish group, namely a more efficient inhibitory control strategy for the speakers of the less typologically similar languages. Within the framework of the DMC model (Braver, 2012) and the application of proactive control strategies during this task (Braver, 2012; Gonthier et al., 2016), this implies that Dutch-Spanish speakers were more effective at employing a proactive control strategy, as reflected in overall shorter RTs. In other words, speakers of typologically more dissimilar languages were better at monitoring and actively maintaining goal-related information compared to speakers of typologically similar languages. This has critical implications for the conceptualisation of the underlying cognitive mechanisms for typologically similar vs. dissimilar language combinations.

With respect to our second research question, there was a descriptive trend of a smaller Stroop effect for the Dutch-Spanish group compared to the Italian-Spanish group. However, the overall processing advantage of the Dutch-Spanish group over the Italian-Spanish group was not reflected in the size of the Stroop effect. More concretely, we did not find a statistically significant difference between the Stroop effect size for the Italian-Spanish group compared to the Dutch-Spanish group. This finding was somewhat surprising and contrasts with our original predictions. Our result suggested, first, that the Stroop effect was unaffected by typological similarity, and second, that speakers of both groups demonstrated a highly comparable inhibitory control performance in this task. Importantly, the CRM framework proposed by Stocco et al. (2014) does not fully account for these specific findings. Instead, our findings strongly suggest a limited modulatory role of typological similarity on inhibitory control performance in this study. One arising question here is the following: why were the Dutch-Spanish speaker faster, but not better, compared to the Italian-Spanish speakers at performing the Stroop task?

One interpretation of our findings could be that factors other than typological similarity influence inhibitory control performance in this task. These other potentially modulating factors exert their

influence such that one group had an advantage in terms of processing speed, but not in terms of overall performance. A well-established modulatory factor in language control, but less in inhibitory control, is language proficiency, as postulated in the IC model (D. W. Green, 1998). Previous studies have shown that multilingual children with a low non-native proficiency display unilateral cross-language interactions from the L1 into the L2 compared to multilingual children with high non-native proficiency (Brenders, Van Hell & Dijkstra, 2011; Poarch & Van Hell, 2012a). As outlined in Poarch and Van Hell (2012b), this could indicate that less language control effort is needed to manage the native and the non-native languages. In turn, this implies less training of more general executive control functions such as inhibitory control if the difference in proficiency levels between the native and non-native language is considerable. More specific to our intermediate late learners of Spanish, one could argue that our participants have not yet sufficiently trained their inhibitory control skills given their intermediate level of non-native proficiency, in turn accounting for a limited effect of typological similarity in this study. Therefore, one possibility is that there is an interaction effect between typological similarity and non-native proficiency, and only a particular degree of typological similarity paired with a specific proficiency level leads to training of the inhibitory control skills. This tentative hypothesis is partially in line with language control research by Brauer (1998). This study explored the effect of typological similarity on language control via the *within-language Stroop effect* and the *between-language Stroop effect* in speakers of typologically similar languages (German-English) and typologically dissimilar languages (English-Greek and English-Chinese) in the classical Stroop paradigm. The within-language Stroop effect refers to the differences in RTs between the congruent and incongruent condition when the stimulus and response languages are identical. On the other hand, the between-language Stroop effect quantifies the differences in RTs between the congruent and incongruent condition when the stimulus and response languages are different (Brauer, 1998; Marian et al., 2013; Van Heuven et al., 2011). Critically, Brauer (1998) included low- and high-proficient speakers to also explore the effect of proficiency on in-

hibitory control performance. All three groups showed a within-language and a between-language Stroop effect. On the one hand, low proficiency in the non-native language was linked to larger differences between the within-language and the between-language Stroop effect across the native and non-native language, irrespective of typological similarity. On the other hand, highly proficient speakers in the typologically dissimilar group were linked to larger within-language compared to between-language Stroop effects in both the native and non-native language. Importantly, highly proficient speakers in the typologically similar group showed no difference between the within-language and the between-language Stroop effect. Therefore, these results suggest that when the difference in proficiency levels is considerable (i.e., low proficiency in the non-native language), the effect of typological similarity on language control performance may be limited, potentially because the amount of "training" of the inhibitory skills has not yet been sufficient to elicit any typological similarity effects.

Given the strong link between language control and domain-general inhibitory control (Bialystok et al., 2012; Declerck et al., 2021; Festman et al., 2010), this argument could be applied to our study: our Italian-Spanish and Dutch-Spanish speakers were late language learners of Spanish who had a B1/B2 proficiency level in Spanish. We therefore postulate that the difference in proficiency between the native language (i.e., Italian or Dutch) and the non-native language Spanish was too substantial to elicit a typological similarity effect on inhibitory control performance, even at intermediate B1/B2 proficiency levels. However, we anticipate that with increasing non-native proficiency levels, a typological similarity effect on inhibitory control may be more pronounced. In view of this, it may not be surprising that inhibitory control performance (i.e., the size of the Stroop effect) was statistically equal given that our groups had highly comparable proficiency levels in their non-native language Spanish. Thus, while our findings are not fully compatible with the CRM framework proposed in the introduction (Stocco et al., 2014), they suggest that at intermediate non-native proficiency levels, the modulating role of typological similarity is not yet trace-

able at the behavioural level.

A second interpretation of our results could be that management of cross-language interference between two typologically similar languages does not directly transfer to strengthening the networks underlying inhibitory control. While we know that speaking multiple languages has a direct impact on language control (Coderre et al., 2013; Coderre & Van Heuven, 2014; D. W. Green, 1998; D. W. Green & Abutalebi, 2013; Mosca & De Bot, 2017), this may not generalise to broader executive functions such as inhibitory control. Contrary to the predictions by the CRM (Stocco et al., 2014), it may be the case that speaking typologically similar languages does not result in a quantitative difference in the amount of training of executive functions over time compared to typologically dissimilar languages. Therefore, the link between speaking typologically similar languages, language control and inhibitory control needs to be more closely inspected in future studies, specifically, the association between language control and inhibitory control.

Considering our compelling findings, the current study takes an important step towards understanding the relative contribution of typological similarity to inhibitory control performance. Taken together, our results suggest that typological similarity only plays a limited role in modulating inhibitory control performance, already at the stage when there is a moderate difference in proficiency levels between the native and the non-native language. However, typological similarity may start to play a role only when non-native proficiency becomes more native-like. Second, our findings further suggest a more complex link between managing multiple languages and more general inhibitory control skills. This could imply that multilingualism primarily influences language control, but that it has only limited effect on domain-general inhibitory control mechanisms. Therefore, our results have important implications for the conceptualisation of the underlying processes of inhibitory control and add novel evidence to the debate around the role of typological similarity in inhibitory control performance.

7.4.1   Conclusions

In this study, we used a spatial Stroop task to examine whether and how inhibitory control performance measured via the Stroop effect was modulated by typological similarity. We found that the typologically dissimilar (Dutch-Spanish) group was faster in performing the task compared to the typologically similar (Italian-Spanish) group. This implied that the Dutch-Spanish group was better at monitoring goal-related information throughout the task compared to the Italian-Spanish group. Critically, this did not impact the overall Stroop task performance. Instead, the size of the Stroop effect, and in turn inhibitory control performance, were similar across both groups, irrespective of typological similarity. Therefore, our results suggest that typological similarity plays a limited role in modulating inhibitory control performance, particularly in intermediate proficient multilinguals with considerable differences in proficiency between their L1 and non-native language(s).

7.4.2   Future directions

Our findings open new avenues to expand on current theoretical frameworks describing the impact of typological similarity on inhibitory control. An emerging line of research could focus on quantifying the degree of interference between typologically similar vs. dissimilar languages and the consequences for language control and/or inhibitory control. For this, future studies should investigate first, language pairs with varying degrees of typological similarity, second, include separate measures for both language control and inhibitory control performance and, finally, recruit speakers of different proficiency levels to tease apart the potentially critical effects of proficiency in modulating inhibitory control performance. Recent years have also seen an increase in research on the neuro-cognition of inhibitory control which combines behavioural measures with electrophysiological and neuroimaging methods (Abutalebi et al., 2012; Christoffels et al., 2007; Constantinidis & Luna, 2019; Grundy, Anderson & Bialystok, 2017). Future studies in this area of research should also incorporate both offline and online measures such as

electroencephalography or fMRI measures to model the cognitive and neural mechanisms underlying inhibitory control performance in multilingual language processing.

CRediT author contribution statement

Sarah Von Grebmer Zu Wolfsthurn: Conceptualisation, Methodology, Validation, Investigation, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review and Editing, Visualisation. **Anna Gupta**: Conceptualisation, Investigation, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review & Editing. **Leticia Pablos-Robles**: Conceptualisation, Methodology, Writing-Review and Editing, Supervision. **Niels O. Schiller**: Conceptualisation, Writing Review and Editing, Supervision, Funding Acquisition.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported here.

Acknowledgements

Funding statement

Data availability statement

The data that support the findings of this study are openly available in the Open Science Framework at: `https://osf.io/e5ba9/?view_only=e8c7dbdf07984cbebb0fce50a132a605`

Citation diversity statement

Within academia, research is witnessing a systematic under-representation of female researchers and members of minorities in published articles (Dworkin et al., 2020; Rust & Mehrpour, 2020; Torres et al., 2020; Zurn et al., 2020). With this Citation Diversity Statement, we aim to raise awareness about this issue. We classified the first and last author based on their preferred gender for each reference in our reference list (wherever this information was available). Our reference list contained 25% woman/woman authors, 42% man/man, 13% woman/man and finally, 16% man/woman authors. For lack of direct comparison in the psycholinguistic field, we compared this to 6.7% for woman/woman, 58.4% for man/man, 25.5% woman/man, and lastly, 9.4% for man/woman authored references for the field of neuroscience (Dworkin et al., 2020). Note that the limitations of this classification are twofold: first, we need to develop adequate comparison metrics for every research field; and second, we need to improve this rudimentary binary gender classification system. However, we are confident that future work will address both issues.

Appendix

7.A   Linguistic profile: Italian-Spanish group

Table 7.A.1: *Linguistic profile of the Italian-Spanish group (N = 33) according to the LEAP-Q (Marian et al., 2007).*

	L1	L2	L3	L4	L5	Total
Italian	n = 33					**33**
Spanish		n = 2	n = 18	n = 10	n = 3	**33**
English		n = 27	n = 5			**32**
French		n = 4	n = 8	n = 3		**15**
German			n = 1	n = 2		**3**
Catalan				n = 1	n = 1	**2**
Portuguese					n = 3	**3**
Total	**33**	**33**	**32**	**16**	**7**	

7.B   Linguistic profile: Dutch-Spanish group

Table 7.B.1: *Linguistic profile of the Dutch-Spanish group (N = 25) according to the LEAP-Q (Marian et al., 2007).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Dutch	n = 25					**25**
Spanish			n = 9	n = 9	n = 7	**25**
English		n = 23	n = 2			**25**
German			n = 7	n = 7		**14**
French		n = 1	n = 7			**8**
Portuguese				n = 2	n = 2	**4**
Frisian		n = 1				**1**
Japanese				n = 1	n = 1	**2**
Italian					n = 1	**1**
Mandarin					n = 1	**1**
Total	**25**	**25**	**25**	**19**	**12**	

7.C Model parameters: Italian-Spanish group

Table 7.C.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for the Italian-Spanish group (n = 32).*

	Formula: accuracy ~ condition (congruent vs. incongruent) + (1|subject) + (1|item)			**Formula:** RTs ~ condition (congruent vs. incongruent) + (condition|subject) + (1|item)		
Term	**Odds ratio [95% CI]**	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	24.86 [16.56 - 37.32]	15.50	< 0.001	579.21 561.94 - 596.48]	65.77	< 0.001
Condition [incongruent]	0.580 [0.400 - 0.784]	-3.38	**0.001**	29.46 [18.10 - 40.82]	5.09	**<0.001**
Random effects						
σ^2	3.29			10783.88		
$\tau_{00Subject}$	0.72			2102.95		
τ_{00Item}	0.01			8.82		
$\tau_{11Subject[incongr.]}$				307.02		
$\rho_{01Subject[incongr.]}$				-0.28		
ICC	0.18			0.16		
$N_{Subject}$	32			32		
N_{Item}	4			4		
Observations	3072			2856		
Marginal R^2 / Conditional R^2	0.020/0.198			0.017/0.173		

7.D Model parameters: Dutch-Spanish group

Table 7.D.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for the Dutch-Spanish group (n = 25).*

Term	**Formula**: accuracy ~ condition (congruent vs. incongruent) + (condition|subject)			**Formula**: RTs ~ condition (congruent vs. incongruent) + (condition|subject)		
	Odds ratio [95% CI]	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	25.65 [17.30, 38.02]	16.16	< 0.001	559.80 [539.98, 579.63]	55.37	< 0.001
Condition [incongruent]	0.477 [0.295, 0.771]	-3.02	**0.003**	15.31 [4.40, 26.21]	2.75	**0.006**
Random effects						
σ^2	3.29			12180.85		
$\tau_{00\,Subject}$	0.33			2289.48		
$\tau_{11\,Subject[incongr.]}$	0.43			226.95		
$\rho_{01\,Subject[incongr.]}$	-0.20			-0.78		
ICC	0.12			0.13		
$N_{Subject}$	25			25		
Observations	2,400			2,236		
Marginal R^2 / Conditional R^2	0.035/0.155			0.004/0.136		

7.E Model parameters: Stroop effect

Table 7.E.1: *Comparison between the model with the interaction effect of condition and typological similarity (left) and the best-fitting model (right) with main effects for condition and typological similarity (n = 57).*

Term	**Formula:** RTs ~ typological similarity TS (high vs. low) * condition (congruent vs. incongruent) + (condition|subject) + (1|item)			**Formula:** RTs ~ typological similarity TS (high vs. low) + condition (congruent vs. incongruent) + (condition|subject) + (1|item)		
	Estimate [95% CI]	t-value	p-value	Estimate [95% CI]	t-value	p-value
(Intercept)	559.79 [540.73, 578.85]	57.57	< 0.001	553.48 [534.99, 571.96]	58.68	< 0.001
TS [high]	19.44 [-5.93, 44.82]	1.50	0.133	30.70 [7.72, 53.68]	2.62	**0.009**
Condition [incongruent]	15.33 [4.36, 26.29]	2.74	**0.006**	23.20 [14.56, 31.83]	5.27	**<0.001**
TS [high] * Condition [incongruent]	13.98 [-0.396, 28.36]	1.91	0.057			
Random effects						
σ^2	11399.52			11397.84		
$\tau_{00\,Subject}$	2101.67			2210.02		
$\tau_{00\,Item}$	1.10			5.03		
$\tau_{11\,Subject[incongr.]}$	243.08			307.02		
$\rho_{01\,Subject[incongr.]}$	-0.50			-0.50		
ICC	0.14			0.15		
$N_{Subject}$	57			57		
N_{Item}	4			4		
Observations	5,092			5,092		
Marginal R^2 / Cond. R^2	0.023/0.161			0.027/0.170		