THE EFFECT OF NOVEL LINGUISTIC MAPPINGS ON INHIBTORY CONTROL

Nuala Harvey
Linguistics, 2021

Abstract: This piece of work endeavoured to analyse the link between executive function and linguistic mapping with the view that, because of shared processes, proficiency in one would improve the other. This was done by use of a preliminary Stroop task to measure participants' initial inhibitory control strength, an artificial language learning task wherein participants were trained on either a holistic or compositional language, and a second Stroop task to measure any temporary enhancement in inhibitory control caused by the artificial language learning. Due to issues with the construction of the Stroop task, the predicted decrease in Stroop effect between the first and second Stroop tasks could not be found. The majority of participants became quicker by the second Stroop task but whether this result was caused directly by the linguistic mapping task remains to be seen, as it could have easily been due to their practice with the first task. That is not to say that the hypothesis was falsified, only that the Stroop task would need redesigning if the experiment were to be carried out again. A secondary aim of the study was to make a point about learnability wherein structured, compositional languages were expected to be easier to learn than fully holistic ones, hence the two variations of the artificial language task. Participants trained on the compositional language were expected to record a more decreased Stroop effect than those who learned the holistic. While the Stroop effect measure has not worked, participants that learned the compositional language did so more faithfully and with more success than the holistic group, supporting this hypothesis.

Keywords: inhibitory control, Stroop task, artificial language learning, bilingualism, executive function

Supervisor(s): Dr Christine Cuskley and Dr Joel Wallenberg

1. Introduction

The link between linguistic ability and executive function has been under investigation since before the conception of modern cognitive science. Articles dating back as far as 1934 and beyond speculate around how '[e]xecutive abilities […] obviously may imply the ability to speak effectively' (Pear 1934: 57). Executive function is the set of cognitive abilities present in the prefrontal cortex that help people with processes such as 'attention, emotion regulation, flexibility, inhibitory control, initiation, organization, planning, self-monitoring, and working memory' (Goldstein et al. 2014: 4), all of which are used in day-to-day life. This dissertation aimed to measure the effects of novel linguistic mappings on just one component of executive function: inhibitory control.

Inhibitory control, in the words of Isquith et al. (2014: 335), is 'the ability to withhold or defer responding to stimuli whether they are internal to the person, such as task-irrelevant impulses, or external such as distractions in the environment.' This is necessary for something as basic as having a conversation in a room with music playing or other people also having their own conversations. Inhibitory control and only inhibitory control was chosen so as to have a specific process to target the investigation towards rather than tackling all of the processes that make up executive function. The strength of it can also be measured in a fairly straightforward manner through the use of a Stroop task with there being a wealth of literature on the connection between the two.

A lot of the Stroop and inhibitory control literature also refers to the connection between bilingualism and enhanced executive function. Following on from this, this study aimed to simulate the processes that apparently lead to this improved executive function in bilingualism through the use of an artificial language learning (ALL) task, preceded and succeeded by Stroop tasks. The Stroop tasks served to establish a baseline value for each participants' inhibitory control, and record the difference in inhibitory control (if any) after performing one of two ALL tasks.

A secondary aim of this study was to investigate whether a more learnable language led to participants' inhibitory control being enhanced even further temporarily. Of the two ALL tasks, one of the artificial languages was designed to be more learnable than the other, given that it was structured and compositional. The other language was completely holistic.

The hypotheses for this study were as follows:

1. Participants that learned the artificial language more effectively, whichever one it was, would record a decreased Stroop effect on the second Stroop task in comparison to their first one. The results for this hypothesis were inconclusive due to an issue with the construction of the Stroop task that will become apparent throughout the rest of this work.
2. The participants that learned the compositional artificial language would have a bigger decrease in Stroop effect than those that learnt the holistic language given that the compositional language was hypothesised to be more learnable. Although the Stroop effect measure itself fell through, this hypothesis was partially supported in that the participants that learnt the compositional language were much more successful in the ALL task test blocks than the holistic language group. The results from this task support the hypothesis that more compositional languages are more learnable and more holistic languages are less so.
3. The final hypothesis was that bilinguals would outperform monolinguals in every area. Given that this study was based in part on research that claims that bilingualism leads to enhanced executive function, it follows that bilinguals would outperform their monolingual counterparts in every aspect of the study. The results here were, again, inconclusive. In the ALL task, bilinguals often started off better than monolinguals, but then did not improve as much. Reasons for this are expanded on in the Discussion.

The background research that led to these hypotheses is discussed throughout the next section. This section involves an in-depth review of the current literature on the journey of the relationship between bilingualism and executive function, which has undergone quite the transformation in the second half of the twentieth century. There is also information on the origin of the artificial languages used in this study and the reason for their choosing. This leads on to a description of the method of this experiment, results and a discussion thereof.

2. Literature Review and Background

2.1 Bilingualism and Executive Function

The impetus behind this study began with claims made about executive function and bilingualism. Before the cognitive revolution that began in the 1950s, bilingualism was largely believed to be something that would put an individual at a cognitive disadvantage

(Barac and Bialystok 2011). This archaic view has now largely been wiped out by a new school of thought that finds that bilingualism does not only not put an individual at a disadvantage, but actually strengthens many aspects of their cognition.

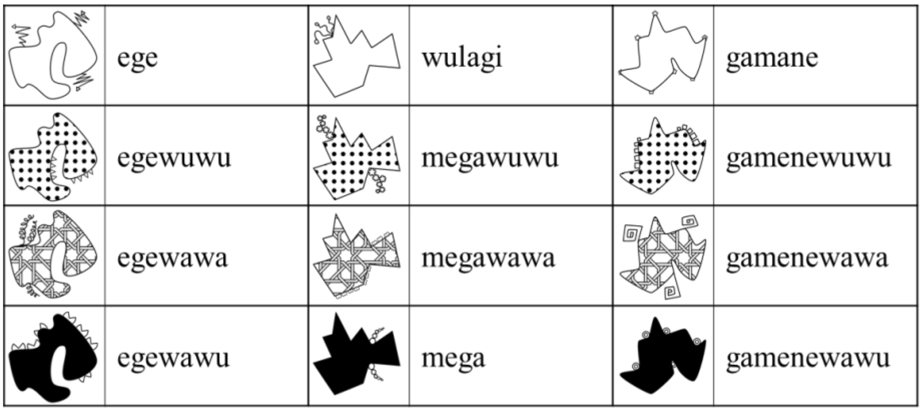
Luk et al. (2011: 588) exemplify this more recent view, finding in their study that 'being actively bilingual is associated with greater advantages in cognitive control and higher language proficiency.' This is just one of many studies done in this area, along with work such as that of Costa et al. (2008), Czapka et al. (2020) and the seminal study of Peal and Lambert (1962). The idea behind this link between bilingualism and executive function is that bilinguals are constantly dealing with two competing systems. Even in a monolingual environment, 'bilingual speakers cannot suppress activation from their first language', shown in the study by Hermans et al. (1998: 213). This means that the different aspects of executive function are being used more frequently in a bilingual brain than a monolingual one, which, in turn, strengthens them. A bilingual individual has to manage 'two representational systems, both rich in detail and structure, that underlie language production' (Bialystok 2007: 210). Even at a basic level, such as having to suppress one or more lexical items in order for the corresponding one in the appropriate language to 'win', this is a significant and constant undertaking – and this does not even cover the more complex ways in which a speaker's syntax, morphology, etc. could be affected. '[A]ttention, inhibition, monitoring and switching' are all 'components of the executive function' as well as being 'the processes necessary to control the two language systems' (Bialystok 2007: 212). Therefore, it follows logically that bilinguals would have some modification or better development in their executive function to accommodate for this additional cognitive load. This is the theory behind the hypothesised improved executive function in bilinguals.

However, in the wealth of studies performed in this area, there have been some unsubstantiated or irreplicable conclusions. Dick et al. (2019) attempted to replicate research that found an executive function bilingual advantage, taking care to account for factors not considered in the original studies, such as 'age, biological sex, race/ethnicity, highest degree of education, household income, marital status, crystallized and fluid intelligence, and English vocabulary' (Dick et al. 2019: 693). This study concludes that 'when [they] properly controlled for covariates, [they] failed to find a bilingual advantage for executive function' (Dick et al. 2019: 695) and suggests that 'previously reported significant effects for executive function in the bilingual literature may reflect type I error' (Dick et al. 2019: 697). Morton (2015: 352) arrives at the same conclusion, finding that some studies interpreted results to be evidence of bilingual advantage, even though the so-called evidence was 'simply introduced by statistical means.' He also, in line with Dick et al. (2019), finds that 'most bilingual advantage research still lacks basic measures of language proficiency and [socioeconomic status], and compares groups of monolinguals and bilinguals that differ in ways beyond language status' (Morton 2015: 353).

This study looked at whether this effect of enhanced executive function in bilinguals could be replicated on a much smaller scale. Participants were given short term exposure to a linguistic mapping task with a miniature 'alien' language and tested for improvement in their inhibitory control via a Stroop task. The aim was to simulate new vocabulary learning in a way that would trigger the same executive function processes that bilinguals are proposed to use constantly. Every time participants were tested on their new vocabulary, they may have been suppressing their own language(s) in order to facilitate the new words. Given this simulation of bilingual vocabulary learning, the hypothesis was that participants' inhibitory control would have been temporarily enhanced thanks to the ALL task, resulting in a reduced Stroop effect on the final Stroop task. Additionally, it was expected that any bilingual participants would perform better across the board compared to monolinguals in line with the research discussed earlier in this section.

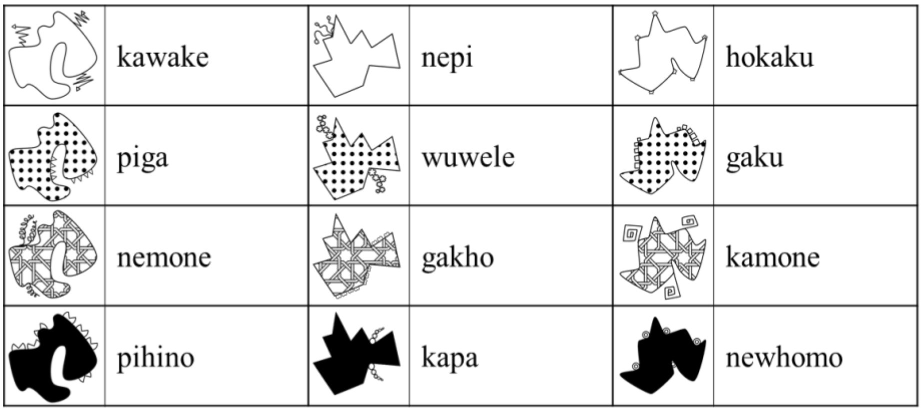
2.2 Structured Linguistic Mappings and the Facilitation of Inhibitory Control

Participants in this study were tested on one of two artificial languages. One is more structured, using fewer strings in different combinations to convey meaning, while the other is fully holistic with a completely separate string for each meaning. These artificial languages are drawn from the end result of Kirby et al.'s (2015) study on the cultural evolution of linguistic structure. In that study, participants were presented with 12 meaning-signal pairs.



	ege		wulagi		gamane
	egewuwu		megawuwu		gamenewuwu
	egewawa		megawawa		gamenewawa
	egewawu		mega		gamenewawu

Figure 1. The original language resulting from the transmission chain in Kirby et al.'s (2015) study, before modification.



	kawake		nepi		hokaku
	piga		wuwele		gaku
	nemone		gakho		kamone
	pihino		kapa		newhomo

Figure 2. The original language resulting from the closed group in Kirby et al.'s (2015) study, before modification.

The initial artificial language was constructed by concatenating 2, 3 or 4 syllables with the structure CV. The resulting signals were then filtered to exclude any items that resembled English words. These languages were then stimuli for two iterative processes. In both versions of the experiment, two participants were trained on an artificial language and then took turns to be Speakers and Hearers. Speakers were presented with a 'topic' (Kirby et al. 2015: 96) which was one of the meanings that they had been trained on. They wrote what they thought the correct signal was and the Hearer had to pick the correct signal out of 6 options which included both the actual topic signal and the Speaker's output. Both were then given feedback and the roles switched. Each participant in the pair performed both roles twice for each meaning. The output from the second time that one of the participants was Speaker was then used as the stimuli for the following iteration of the task. In one version, this new language went to a different pair of participants with each new iteration, forming a six-generation transmission chain. In the other version, the same participants were then retrained with their own output as the new language in a six-generation closed group. One of the transmission chains output the structured language seen in *Figure 1* after the sixth generation. One of the closed groups output the holistic language seen in *Figure 2* after the sixth generation. The images shown in the tables are different from those used in the original study for reasons discussed in §3.2.2.

In a previous study of a similar nature, Kirby et al. (2008: 10682) hypothesise that 'we would see the emergence of adaptive structure in response to the pressure on the language to be transmitted faithfully from generation to generation.' The language of the closed group was only under a pressure for expressivity. Given the closed group's common ground that grew as the experiment progressed, their resulting language did not have the same pressure for compressibility that the transmission chain's language did. Participants in the closed group were more familiar with each other as well as the meanings themselves. Say, for example, that one participant in the closed group frequently misremembers the name for one of the meanings over multiple trials. The fact that both of these participants have performed this task before and will be familiar with each other's mistakes means there is less pressure for compressibility. One participant could even make the same error so many times that both speakers adopt that signal over the original. In the other group, this common ground is not established as each pair performs the task once. Participants responded to the pressure of the transmission chain by creating a more learnable language and more learnable, in this case, meant more structured. The combination of pressures for expressivity and compressibility is what gave rise to the structure observed in the language in *Figure 1*. It can be hypothesised that 'the reduction in the number of strings must make a language easier for participants to learn' (Kirby et al. 2008: 10683).

Given this hypothesis, participants that learned the more structured, compositional language in this experiment were expected to achieve a higher level of accuracy in the test blocks than participants that learned the holistic language. Following from this, it was predicted that participants that learned the artificial language with the most success would also record the most decreased Stroop effect in the final task. The better that a participant learned the artificial language, whichever one it was, the more likely it was that they would be simulating an aspect of the bilingual process which should strengthen their inhibitory control. Participants that learned the structured language were expected to have a greater decrease in Stroop effect by the end of the study given that that was the more learnable language. All participants were expected to have a small decrease in Stroop effect in correlation with familiarity with the task, given that it was the second time they were performing it in the space of 30 minutes.

3. Methods

3.1 Participants

Participants were sourced through Dr. Christine Cuskley's and my own social media accounts. Participation was voluntary. The information sheet that participants were presented with at the beginning can be found in the Appendix along with the debrief that participants were given upon completion of the experiment. In total, there were 50 participants: 32 women, 16 men and 2 non-binary individuals. The options for gender in the questionnaire were limited to 'man', 'woman' or 'other'. It is important to consider whether the two non-binary participants should be grouped together given that non-binary is an umbrella term that covers a number of gender identities (Stonewall 2020). In fact, they may not have been any more similar to each other than they were to the other groups and, were gender differences the focus of this study, a more detailed description of their gender identities would have been collected. The youngest participant was 18 and the oldest was 58, with approximately two thirds of participants aged 20 to 23. The median age of the group was 22. The mean age of the group was 24.6. Forty-three participants were L1 English speakers with the remaining 7 participants acquiring English from the ages of 4 to 10. Within the L1 English speakers, 19 reported speaking another language as well. However, participants were not asked to quantify the degree to which they spoke this other language so these participants may have varied from having some L2 knowledge all the way to balanced bilingualism. Given that bilingualism was hypothesised to assist participants in this experiment, the 7 L2 English speakers could be used as a more reliable group with which to test this hypothesis given that they were at least proficient enough in their L2 to be able to partake in this experiment in English. Two participants reported colour-blindness which may have negatively impacted their results on the Stroop tasks which will be investigated further in the Results section.

Data regarding pre-existing cognitive disorders was not collected as it was deemed unethical to gather this information. However, it is important to note that there is evidence that some cognitive disorders can negatively impact inhibitory control, meaning that some participants may have returned results that are more to do with their own, personal neurological make-up that could not be extended as results that would reflect those of the neurotypical population. Reiter et al. (2005: 124) find in their study that 'a mild impairment in inhibiting inadequate reactions as well as an increase in mental processing time can be seen in dyslexic children' via the use of a Stroop task. According to Knight (2018: 207), 5 – 10% of the worldwide population is dyslexic with varying severity of the disorder. Therefore, it is not unlikely that some of the participants in this study could have been dyslexic and this was taken into consideration when analysing the results and could be an explanation for anomalous data.

Dyslexia is not the only disorder of this nature, as Attention Deficit Disorder (ADD) also affects an individual's Stroop performance. Stroop tasks have been used to show the effects of ADD many times as ‘children with ADHD are known to have an impaired ability to perform’ on them (Klingberg et al. 2002: 782). Whether an individual has ADD with hyperactivity or not, Barkley et al. (1992: 172) conclude that ‘no differences were found between the subtypes of ADD’ in terms of their Stroop results. Both Barkley et al. (1992: 169) and Kóbor et al. (2015: 344) find that individuals with AD(H)D were slower to respond to Stroop tasks but not necessarily less accurate. Both these articles also reflect on the heterogeneity observed within the AD(H)D population. Therefore, even if information pertaining to cognitive disorders had been collected from this study’s participant group, this still might not have explained the results from any participants that may have had AD(H)D. However, it is still necessary to take this into account during the analysis and discussion of the results collected.

Work by Lezak et al. (2004) has also shown that various types of head trauma or damage to the frontal cortex will have a negative effect on inhibitory control, causing participants to perform worse than the neurotypical population on Stroop tasks. Autism Spectrum Disorder (ASD) is also associated with lower accuracy in Stroop tasks (Robinson et al. 2009). All this is to say that caution was taken when analysing the results of this study as there are many instances where a participant’s neurodiversity could masquerade as a result of the experimental process, rather than being seen correctly as the result of a pre-existing disorder.

MacLeod (1991: 184) remarks that ‘[t]here are no sex differences in Stroop interference’ and that, after the initial onset of the Stroop effect when an individual learns to read, interference does not begin to increase again until ‘approximately age 60’ (MacLeod 1991: 185). Given that all participants in this study were aged 18 to 58, variation between the different ages was not expected, nor was variation between the different genders.

3.2 Materials

3.2.1 Stroop Task

There were two Stroop tasks in this study, one before the ALL task and one after. Each task was made up of 80 trials and there were two different sequences of trials so that participants saw a different order each time they performed the task. These two orders were counterbalanced so that 50% of participants saw one order the first time and the other order the second, while the other 50% saw the opposite.

Constraints were put into place to ensure that neither order was easier than the other and these will be discussed later in this section. However, this further precaution of counterbalancing the orders controlled for any instance where one order facilitated or inhibited a participant's performance in the ALL more than the other. Both these orders can be seen in full in the Appendix.

There were 16 possible name-ink pairings in the Stroop tasks, each shown in *Table 1*. Each pairing appeared five times per task. 80% of the pairings were incongruous, where the name and ink did not match e.g., 'red', with the remaining 20% being congruous e.g., 'red'. This ratio follows on from the Stroop task available from PsyToolkit (2021). As previously mentioned, the sequence of the trials has been pseudorandomised in line with three constraints, two of which are mentioned by C. M. MacLeod in his review of the Stroop effect (1991). They are as follows:

Table 1. Showing all possible name-ink combinations in the Stroop tasks, each appears five times per block.

Colour Name	Ink Colour	Stimulus
Red	Red	red
Red	Blue	red
Red	Green	red
Red	Orange	red
Blue	Red	blue
Blue	Blue	blue
Blue	Green	blue
Blue	Orange	blue
Green	Red	green
Green	Blue	green
Green	Green	green
Green	Orange	green
Orange	Red	orange
Orange	Blue	orange
Orange	Green	orange
Orange	Orange	orange

1. Dalyrmple-Alford and Budayr (1966: 1214) find that 'the structure of the list does have an effect' on the Stroop effect. Specifically, 'the suppression of a response' in the form of colour name 'results in temporary unavailability of that response' in the following trial (Dalrymple-Alford and Budayr 1966: 1213). Say a participant sees the stimulus 'red', they then have to supress the word 'red' in favour of the word 'blue'. If the following trial then consists of the stimulus, 'orange', the residual suppression makes 'red' harder to access which artificially increases the Stroop effect. To avoid this interference, the trials in this study were sequenced in a way that an ink colour was not preceded by its corresponding colour name. In other words, the trials had been sequenced so that the colour name in trial n - 1 ≠ ink colour in trial n.
2. MacLeod (1991: 178) notes that colour naming can be facilitated when the preceding ink colour matches the colour name in the current trial. The idea is that 'having just made a particular response on the last trial makes it easier to discard that as a possible response on this trial.' For this reason, the trial blocks were sequenced in such a way that the ink colour on trial n – 1 ≠ the colour name in trial n.
3. The final constraint is that the exact same combination cannot appear directly before or after itself. Participants would be primed by the trial n – 1 and so their response could be facilitated in trial n. To avoid this effect, trial blocks were sequenced so that ink colour <u>and</u> colour name in trial n – 1 ≠ ink colour <u>and</u> colour name in trial n.

Having reviewed the literature again since generating these constraints, MacLeod (1991: 178) mentions a rule not included here in that a colour name cannot directly follow itself as 'the word is already suppressed and will cause less interference'. This should have been controlled for in this study as there are instances in the trial blocks where this sequencing did occur. However, given the other constraints, the failure to control for this instance did not cause a salient difference in the Results, although there were other factors that became an issue for the Stroop task data, all of which will be discussed in §5.

3.2.2 Artificial Language Learning Task

The stimuli for the artificial language learning task were drawn from Kirby et al.'s study (2015) on cumulative, cultural language evolution. They generated 12 images that consisted of a combination of 'three distinct shapes, and four distinct textures', with each image also having its own 'unique appendage' (Kirby et al. 2015: 95). The exact stimuli could not be used in this study as a high enough quality of image could not be extracted from the original article. However, the images used in this study did still abide by the same parameters of three shapes, four textures and a different appendage for each.

The structured artificial language was modified from the original study to be maximally structured. Given that the original language was the product of a transmission chain and not made with the explicit aim of being structured, there are some anomalies. In this study's version, each free base denoted a shape: 'ege', 'mega' and 'gamene'. Each suffix (or lack thereof) denoted a texture: '-wawa', '-wuwu' and '-wawu'. These rules were applied to all signals without exception. In the original, 'mega' referred to what was called 'megawawu' in this version, and the shape that was in this version called 'mega' was anomalously termed 'wulagi' in the original. There was also some incongruity in the original dataset where the shape in the rightmost column was written 'gamane' as a free base, but 'gamene-' when affixed. This difference was reconciled in this study and all were spelt as 'gamene'. One of the aims of this study was for participants to learn a structured artificial language, therefore, it was necessary to neutralise all incongruous stimuli to make the language as structured as possible.

The holistic language has only undergone one minor modification. The string that was originally 'newhomo' was changed to 'nahomo', to decrease its similarity to existing English words. The original names can be seen in *Figures 1* and *2* and the modified versions in *Figures 3* and *4*.

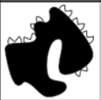
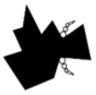
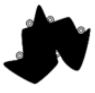
	ege		**mega**		gamene
	egewuwu		megawuwu		gamenewuwu
	egewawa		megawawa		gamenewawa
	egewawu		mega**wawu**		gamenewawu

Figure 3. The modified structured language used in this study with changes highlighted in red.

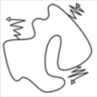
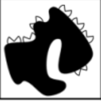
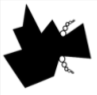
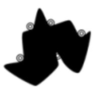
	kawake		nepi		hokaku
	piga		wuwele		gaku
	nemone		gakho		kamone
	pihino		kapa		nahomo

Figure 4. The modified holistic language used in this study with the modification highlighted in red.

3.3 Procedure

3.3.1 Stroop Task

In this study, participants were asked to do an experiment in three parts: an initial Stroop task; an artificial language learning task; and a final Stroop task. The Stroop tasks served to analyse the strength of each participants' inhibitory control as a base line and then measure whether the language tasks had temporarily improved or reduced their inhibitory control.

The Stroop tasks were modelled after the task originally presented by J. R. Stroop in 1935. The theory behind the Stroop effect is that participants have to inhibit the response of an automated cognitive process in order to favour a controlled one. The automated process in the original and this study is reading. The controlled process is the recognition and naming of the colour in which the word is written. When the word in any given trial is a different colour name to the colour of the ink, it should take the participant longer as they must inhibit the written word in favour of the colour name. The need for inhibition is lessened in trials where the colour name and ink colour match. The ‘Stroop effect’ is the disparity in reaction time between trials where the word and colour match and when they do not.

The version presented here is drawn mainly from the example given by PsyToolkit (2021). In any given trial, participants were presented with a white screen with a black cross in the centre for 500ms. One of four colour names, ‘red’, ‘blue’, ‘green’ or ‘orange’, then appeared on the screen in either a congruously coloured ink, e.g., ‘green’, or an incongruously coloured ink named by the other colour names, e.g., ‘green’, ‘green’ or ‘green’. Most traditional Stroop tasks use yellow over orange, but orange was favoured for this study given that it was easier to see against a white background and had no foreseeable drawbacks. Participants were instructed to press the ‘A’ key or ‘L’ key in response to the colour name and ink being a congruous or incongruous match. 50% of participants were instructed to press ‘A’ if the match was congruous and the other 50% were instructed to press ‘A’ if the match was incongruous. This variation controlled for any facilitation that may have been caused by lateralisation. In the PsyToolkit (2021) version, participants have to press the key corresponding to the first letter of the name of the ink colour e.g., ‘R’ for ‘green’, ‘B’ for ‘blue’, and so on. In this version, participants only had to make a judgement on whether the ink colour and word were congruous or not which did not require as much inhibition as naming the ink colour. This caused issues with the Stroop data and is discussed further in §5.

The stimulus remained on screen for 3s. If participants did not respond within 3s, the screen displayed a sad face for 500ms. If the participant responded within the 3s window, the screen displayed a happy or sad face for 500ms as feedback depending on their response. The screen then showed the black cross again and the process repeated. There was an initial demo that consisted of 16 trials to let participants practice. Participants could then go on to their first Stroop task. This task took a maximum of 1m 4s for the demo and 5m 20s per block of 80 trials.

Due to an error in the code, some participants’ data yielded duplicate trials rather than 80 distinct trials. These duplicates have been removed from the results meaning that some of the data is incomplete. However, the data affected only concerned a minority of participants and no block of trials is missing any more than five trials, therefore this should not have a significant effect on the overall results.

3.3.2 Artificial Language Learning Task

After the initial Stroop task, 50% of participants were taken to the holistic ALL task with the remaining 50% being directed towards the structured ALL task. For both tasks, the participants underwent a training block followed by a test, twice. Participants were shown a white screen with a black cross in the centre for 500ms, then each of the 12 meaning-signal pairs appeared twice in a randomised order. Participants saw the meaning first for 500ms, then the corresponding signal for a further 2.5s before returning to the plain screen with the black cross. This was half the amount of time used by Kirby et al. (2015) given that participants in this study saw each pairing twice per training block where they only saw them once in the other study. This training block took 1m 24s.

After the training block, participants were presented with the meaning only and asked to type what they thought the signal was. There was no time limit on this. When they submitted their answer, participants were given feedback. If they answered incorrectly, they were shown what the correct answer was. After they had completed the test, the whole process repeated.

3.4 Analysis

3.4.1 Stroop Analysis

The Stroop results were calculated across a number of measures. Stroop tasks give two sets of results: accuracy and reaction time (RT). The percentage accuracy for each task was calculated for every participant, giving them two each: one for the first task and one for the second. This was done by finding the number of all successful trials, excluding trials where the response was incorrect or timed out, and then dividing this by the total number of trials to give a percentage. One way to measure the hypothesised improvement in inhibitory control is by finding the difference between the percentage accuracy for the first and second tasks. Averages of this difference were found for the participant populations as it was hypothesised that the group trained on the compositional language would show more improvement than the holistic language group.

Each participant has two RT results per task, meaning they had four in total. The average RT in milliseconds (ms) was calculated for all correct congruous trials and then again for all correct incongruous trials with the hypothesis that, in line with the wider Stroop (1935) literature, the incongruous trials would have a longer RT than their congruous counterpart. After the averages had been found, the congruous RT average was subtracted from the incongruous RT average to find the Stroop effect. As with the percentage accuracy, these RT results can then be averaged with those of participants who fall into the same population, whether it be bilingual/monolingual, sex, age, or which ALL task they performed, to find results that can be compared with those hypothesised.

3.4.2 Artificial Language Learning Analysis

The results for the ALL task were calculated using the normalised version of the Damerau-Levenshtein edit distance (Damerau 1964, Levenshtein 1966). Although this measure is being used here to find the distance between two 'words', it has a wide number of applications, such as quantifying the similarity of DNA or protein sequences, spell checking and correcting OCR errors (OpenGenus 2019). This metric assigns a value to the distance between a participant's output and the target word. In this method, there are four types of edits that are permitted: substitution; insertion; deletion; and transposition of two adjacent characters. There are some metrics that disallow some of these operations but, in this version, they are all possible. '[T]he length of the optimal edit sequence is known as the Damerau-Levenshtein (DL) distance' (Zhao and Sahni 2019: 19) as it calculates the fewest number of operations that need to be performed on the output word to reach the target word. For this data, each step between the two words was assigned a value of 1. All these steps were then added together. For example, an edit distance consisting of two substitutions and one transposition would have a value of 3. This value was then divided by the number of characters in the target word to control for the different lengths of target word in this experiment. This final number is what has been used to determine participants' accuracy. An output that perfectly matched the target word would return an edit distance of 0. Any output that required more steps than there were characters in the target word would return an edit distance of 1 or above. The edit distance scores were capped at 1. Given that a score of 1 meant that every single character in the output word was incorrect, it was not considered important to this study how much 'more wrong' a participant was beyond that point.

Before the edit distance was calculated, the data was amended in some respects so as to ensure that the metric could work properly. The code used for the edit distance is case-sensitive, so all responses were put into lowercase to make sure that instances of the right word in the wrong case were not being given a higher edit distance than they should have actually had. There was no ‘skip’ option in the test portion of the ALL experiment. This led to some participants giving null responses when they were not confident enough to attempt the target word. Any responses that were clearly not attempts, such as ‘nope’, ‘I don’t know’, etc., have been removed and replaced with ‘?’, which gives an edit distance of 1. As the ALL data is being used to measure how well participants had learnt the language, null responses have been considered as no attempt at learning the language, which is why they have been assigned a value of 1. This was especially important to amend in cases where participants wrote ‘nope’, given that one of the target words was ‘nepi’. This would give an edit distance of 0.5 making it seem like they were learning. However, it is clear that this was not a legitimate attempt as the participant that did this also wrote ‘nope’ for almost every other trial in that test block.

There are two ways that participants’ success is being judged. The primary measure is the outcome of the second test block. The number of times a participant gave a correct response and the average edit distance value in the second test block are both being used as a proxy for how well participants learnt the artificial languages. The secondary measure is the improvement (if any) between the average edit distance value for the first and second test blocks. This measure is not of as much importance as the first, given that participants who already performed well on the first test block will show little improvement on the second block. This could be misinterpreted as bad performance when, in reality, these participants performed better across the board. To avoid this, the terms of improvement have been normalised in the calculation. This means that an improvement from an initial edit distance of 0.4 to a second score of 0.2 is marked as a 50% improvement, rather than a 20% improvement. While this second measure is interesting, the final scores on the second test block are still being used as the primary method to quantify how well participants learned the artificial languages.

4. Results

4.1 Stroop Tasks

Across all participants, the average % accuracy on the first Stroop task was 97.6%, the lowest accuracy being 91% and the highest being 100%, with a median of 98%. On the second task, this dropped to 96.5% with the range of results growing, the lowest being 81% and the highest being 100% and a median of 97%. However, only two participants, scoring 81% and 84%, fell outside of the range of results seen in the first task.

The average congruous RT for the first task was 753ms with all results ranging from 537ms to 1577ms and a median of 728ms. The average congruous RT for the second task was 684ms with all results ranging from 529ms to 1540ms and a median of 663ms. By mean, range and median, the RT decreased in the second task which is what was expected.

The average incongruous RT for the first task was 747ms with all results ranging from 528ms to 1575ms and a median of 690ms. The average incongruous RT for the second task was 673ms with all results ranging from 484ms to 1560ms and a median of 650ms. As with the congruous results, the RT has decreased across all averages. The graph at *Figure 5* represents this decrease measured in both conditions (congruous and incongruous) across the entire participant population.

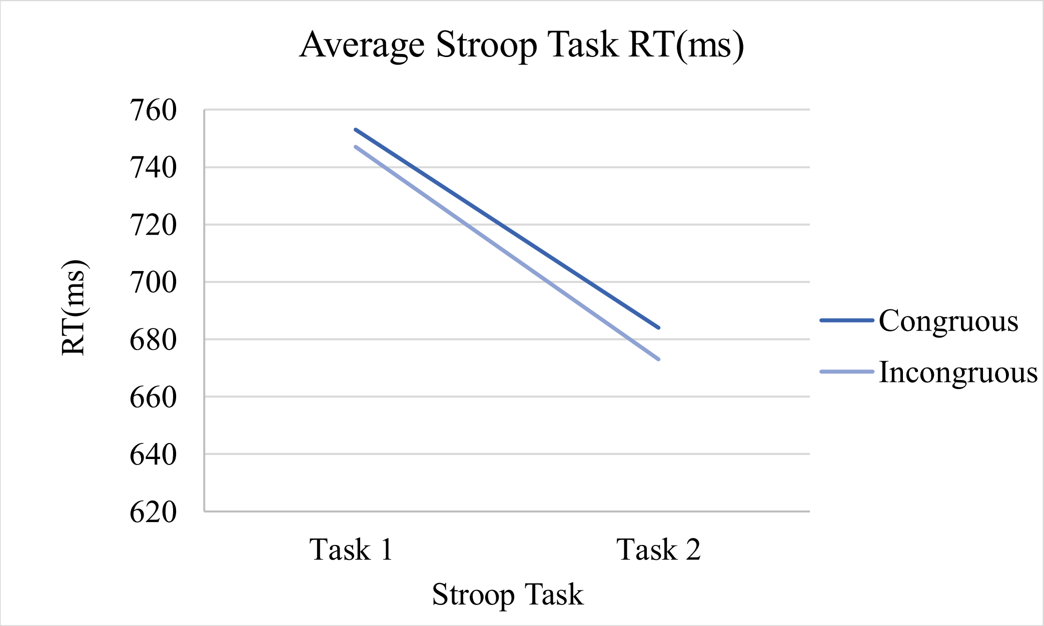


Figure 5 A line graph showing the average RT(ms) of all 50 participants across the two Stroop tasks

If the mean averages presented in this section are used, the first task presents an average Stroop effect of -5ms and -11ms for the second Stroop task. More than half of the participants in the first task recorded a negative Stroop effect and even more were found in the second task. This result is unexpected and will be analysed further in the Discussion. An average for the Stroop effect itself has not been created given that the Stroop effect ranged from -168ms to 419ms in the first task and from -149ms to 149ms in the second. The medians were -10ms and -13ms respectively. These statistics are unreliable for the purpose for reasons to be discussed and so will not play a huge role in our overall results.

The results can also be shown by participant population. The group that were trained on the compositional language had an average accuracy of 97.8% in the first task with this dropping to 95.9% in the second task. The holistic group also saw a drop in accuracy although this was much smaller from 97.4% to 97%. The biggest drop in accuracy over the entire participant group was a drop of 15% from one of the participants in the compositional group, followed by a 13% drop from a member of the holistic group. Both groups, unprecedentedly, consistently took longer to respond to congruous stimuli than incongruous. In the first task, the compositional group averaged a 726ms RT for congruous stimuli and 719ms to respond to incongruous stimuli. In the holistic group, the difference was similar, taking 774ms with congruous stimuli and 769ms with incongruous stimuli. For the holistic group, this gap remained consistent in the second task, taking 696ms with congruous stimuli and 692ms with incongruous stimuli. The gap increased with the compositional group, who took 668ms with congruous stimuli and 650ms with their incongruous counterpart in the second task.

It was predicted that the group that learnt the compositional artificial language would perform better than the holistic language learners on the second Stroop task. However, the compositional group already had a lower average RT in the initial Stroop task than the holistic group, being almost 50ms faster with both congruous and incongruous stimuli. For this reason, a measure of normalised improvement, similar to the one used in the ALL task analysis, gives a fairer representation of the difference between the two groups. If the average RT for the second task were to be compared across the two groups, this could just be capturing a difference between the participant groups not caused by the ALL task, and rather caused by an external factor to the study, like bilingualism. If, instead, the participants' improvement is compared, this is a more reliable measure. Comparing participant populations against themselves, the holistic group improved by 9% across both congruous and incongruous RT, knocking 78ms and 77ms off their average time respectively. Although the compositional group had a faster baseline RT, they only improved by 7% (58ms) on the congruous stimuli and 8% (69ms) on the incongruous stimuli. While this may not seem like a bad result, given that the holistic group only improved by 1% – 2% more, this is all within the context that the compositional group were expected to do better in this task. *Figures 6* and *7* show a comparison for the decrease in RT between the holistic and the compositional groups, the former showing the decrease of the congruous RT between Task 1 and 2 and the latter showing the incongruous version.

The holistic task was performed by 28 of the total 50 participants, 13 of which recorded a standard Stroop effect on the first task, with the remaining 15 all recording negative Stroop tasks. Of these 13, the Stroop effect ranged from 7ms to 153ms, averaging 50ms with a median of 44ms. 7 of these participants went on to record a negative Stroop effect in the

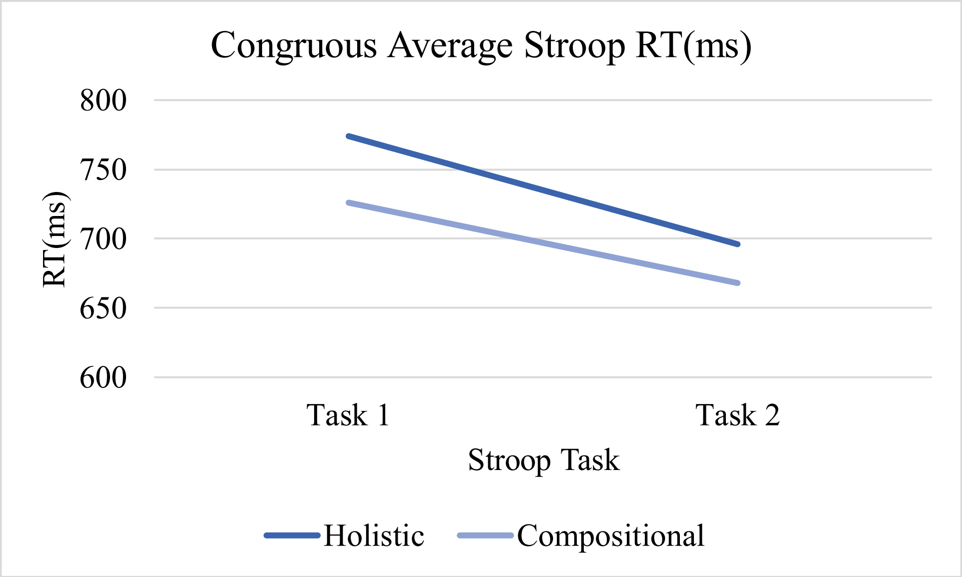


Figure 6 A line graph showing the decrease in congruous stimuli RT(ms)

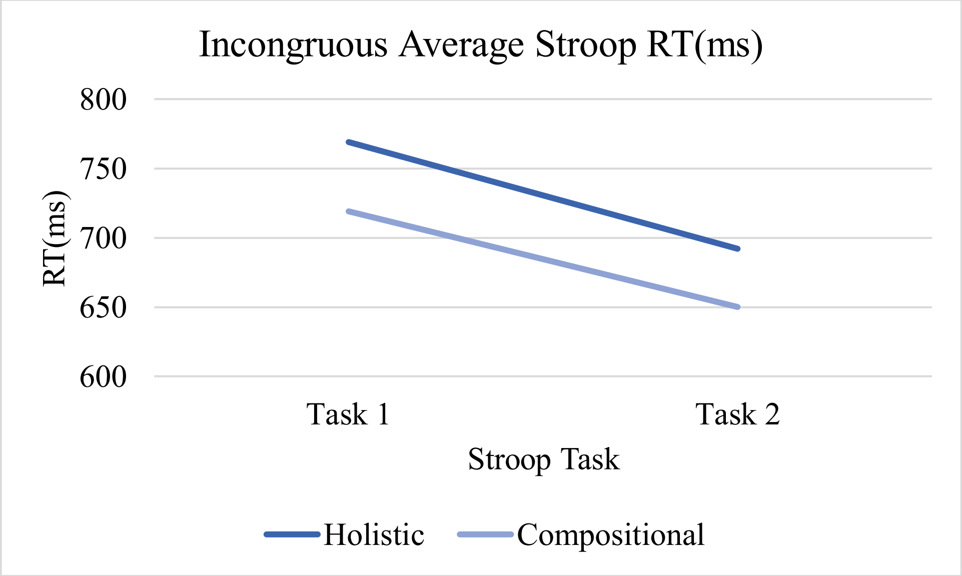


Figure 7 A line graph showing the decrease in incongruous stimuli RT(ms) between the first and second Stroop tasks.

second task, with 4 of the remaining 6 recording an increased or the same Stroop effect. Of the 28 participants in the holistic group, only 2 recorded the expected result of a decreased yet not negative Stroop effect.

Of the 22 participants that completed the compositional task, only 7 recorded a positive Stroop effect on the first task. These Stroop effects range from 2ms to 419ms, averaging 92ms with a median of 29ms. Of these 7, 4 went on to record a negative Stroop effect in the second task. The remaining 3 did all see decrease in Stroop effect in the second task.

Another pertinent way in which to split the population is by mono-/multilingualism. The population is split three ways between English monolinguals, L1 English bilinguals and L2 English bilinguals. As discussed in the participants section, the initial questionnaire did not gauge bilingual proficiency in those that responded that they could speak another language. Any salient difference between the L1 English bilinguals and their L2 English counterparts is likely down to the fact that the L2 English group must have been proficient enough in English to understand and partake in the experiment. However, no such pressure was put on the L1 English bilinguals so their bilingual proficiency could vary a great deal more.

In the first Stroop task, the monolingual group recorded the exact same RT (755ms) as the L1 English Bilinguals for incongruous stimuli, whereas the L2 English bilinguals were 54ms faster. For the congruous stimuli, however, the L2 English bilinguals were actually slowest in the first task (765ms) with the monolinguals recording almost the exact same RT as for the incongruous stimuli (754ms) and the L1 English bilinguals coming in fastest at 746ms. For both congruous and incongruous stimuli, the L2 English speakers decreased their RT the least, lowering it by 56ms and 30ms respectively. The L1 English bilinguals decreased their RT the most, being 77ms faster on congruous stimuli and 91ms faster on the incongruous trials. These results are represented in *Figures 8* and *9*.

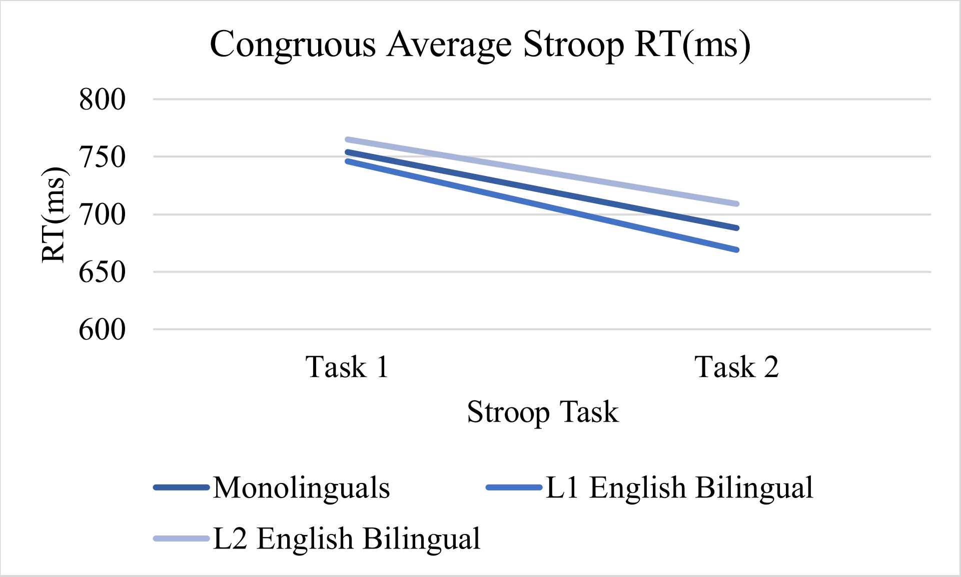


Figure 8 A line graph showing the decrease in congruous RT(ms), split by mono-/multilingual participant populations.

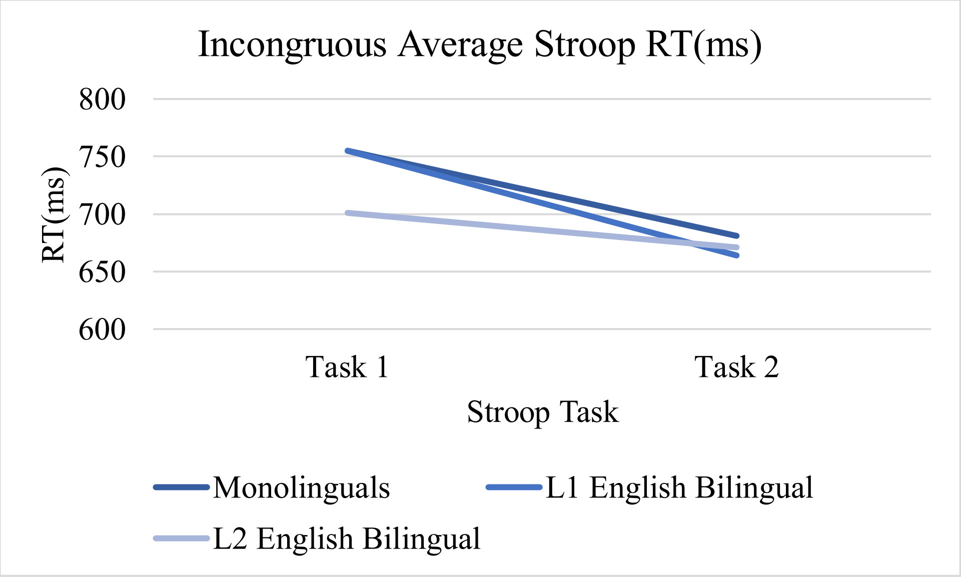


Figure 9 A line graph showing the decrease in incongruous RT(ms), split by mono-/multilingual participant populations.

Of the two participants that identified themselves as colour-blind, one did record RTs above the average in the first Stroop task. However, although above average, there were participants with standard colour vision who recorded higher RTs. For this reason, the data is not being excluded because this result has occurred naturally elsewhere and so cannot be definitively said to be caused by the colour-blindness. By the second Stroop task, this participant's RTs were in line with the rest of the group and the other colour-blind participant did not record any usual results at all.

There are many more ways that the results could be compared, such as contrasting the sexes or age groups. However, for the sake of brevity and relevance, the holistic/compositional and monolingual/bilingual splits are the only ones discussed here.

4.2 Artificial Language Learning Tasks

The results observed in the ALL tasks do support the hypothesis that, in line with Kirby et al. (2008), compositional, structured languages are more learnable. First, to look at the general participant population, participants had an average of 32% accuracy in the second test block. These ranged from 0% to 100% with a median of 25%. The average Damerau-Levenshtein score was 0.41, a median of 0.43 with results ranging from 0 to 0.88. The final measure used to quantify these results was the normalised improvement. Over the whole population, participants improved on average by 23%. The range of all the improvement values fell between -58%, since some participants actually got worse on the second test, and 75%. The median was 24%. As this section continues, the different participant groups will be teased apart so as to clarify these results.

The holistic group's average percentage accuracy in the second block was 20%. The full range of accuracy in the holistic group went from 0% to 58% with a median of 17%. In the compositional group, the average percentage accuracy on the second test block was 47%, over double that of the holistic group. The range of accuracy for this measure in the compositional group went from 8% up to 100% with a median of 38%. This clearly shows that the compositional language was learnt, on average, with more success than the holistic language.

This can be shown with other measures too, such as the average edit distance score for the second block, bearing in mind that a lower score is a better score for this metric. The average and median edit distance score for the second block of trials in the holistic group was 0.58, ranging 0.28 to 0.88. The same figure for the compositional group was just under a third of this at 0.19, ranging from 0 to 0.45 with a media of 0.2.

The tertiary measure for showing this effect is the normalised improvement score. On average, the holistic group improved by 17% between their first and second test block. The compositional group improved by almost double this, on average, making a 30% improvement between tests. The improvement of both groups can be seen in *Figure 10*.

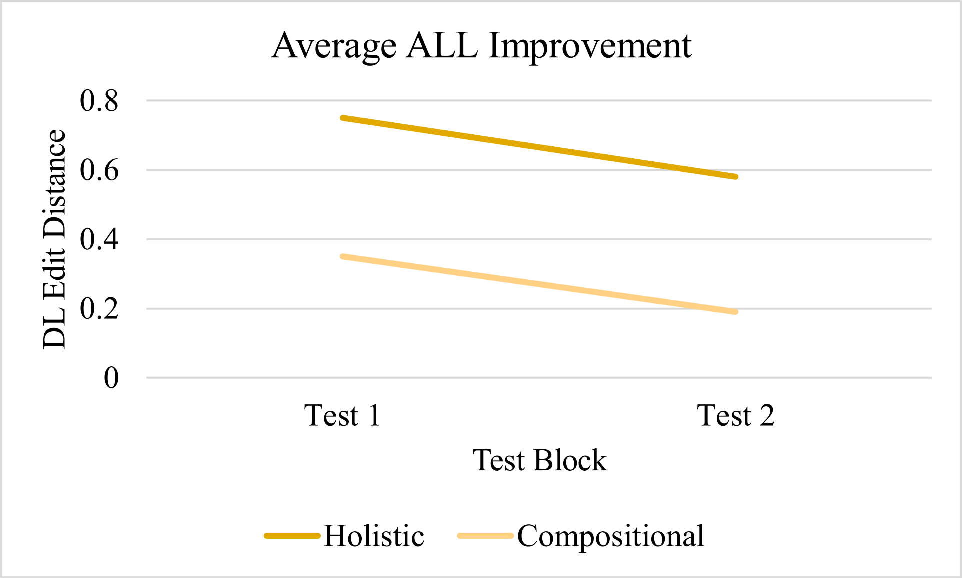


Figure 10 A line graph showing the average improvement of the holistic and compositional groups on the ALL task.

As with the Stroop tasks, the ALL results can also be further broken down by participant group. First, the results can be compared from different groups within the same condition: holistic or compositional. *Figures 11* and *12* show the comparison between the monolinguals, L1 English bilinguals and L2 English bilinguals within the same condition. The bilinguals were expected to learn the artificial language best given that 'bilinguals have advantages over monolinguals in [third language acquisition]' (Cenoz 2013: 75). Given that the L2 English bilinguals are expected to have a consistently higher second-language proficiency, they were expected to outperform the L1 English bilinguals. Surprisingly, the L2 English improved the least in the holistic group with two of the three participants actually performing worse on the second task. The monolingual and L1 English bilingual groups both improved by 20%.

The compositional group performed quite consistently across the board. The L1 English bilingual group performed the best here, showing a normalised improvement of 44% with the L2 English speakers improving by 30% and the monolinguals by 22%.

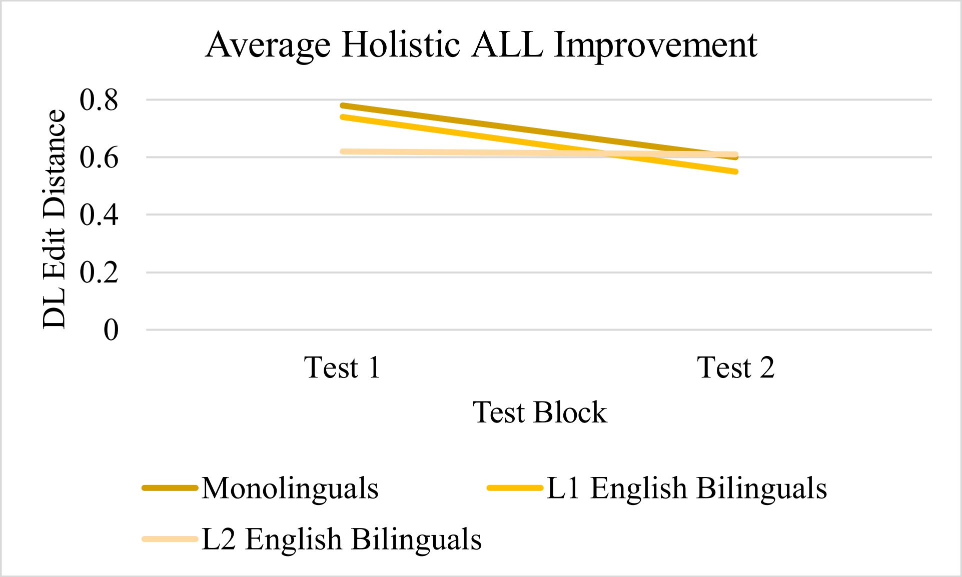


Figure 11 A line graph showing the average improvement for the holistic group on the ALL task.

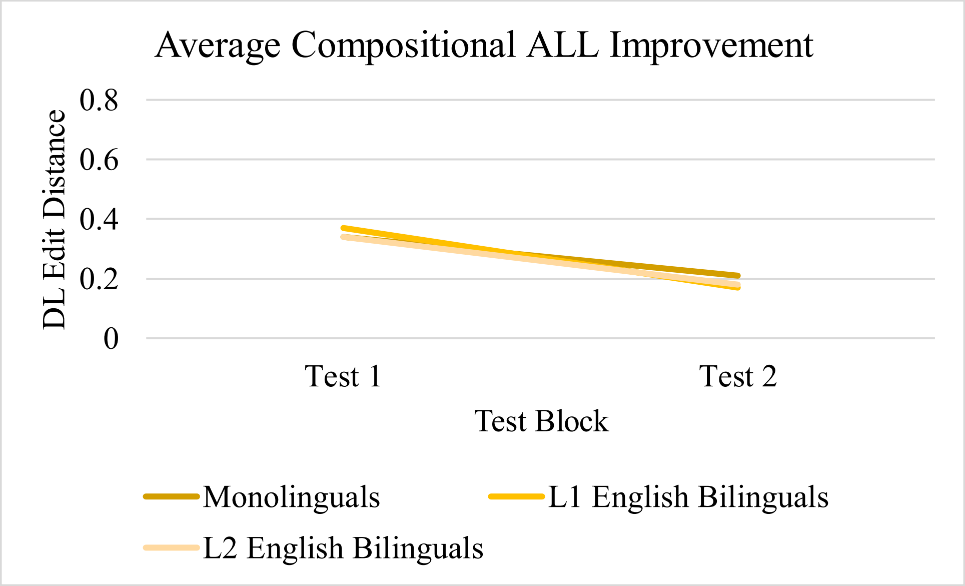


Figure 12 A line graph showing the average improvement for the compositional group on the ALL task.

Figures 13, *14* and *15* show the comparison across conditions. This represents the success of the compositional language learning better as the compositional learners outperformed the holistic group at every opportunity. The graphs may make it seem like the holistic groups had a steeper improvement than the compositional groups, but it is important to bear in mind that the compositional groups had a much lower starting point, so had less room to improve. For the monolinguals, the compositional group's score for the first test was less than half that of the holistic group and reduced further to just over a third of the holistic score by the second test.

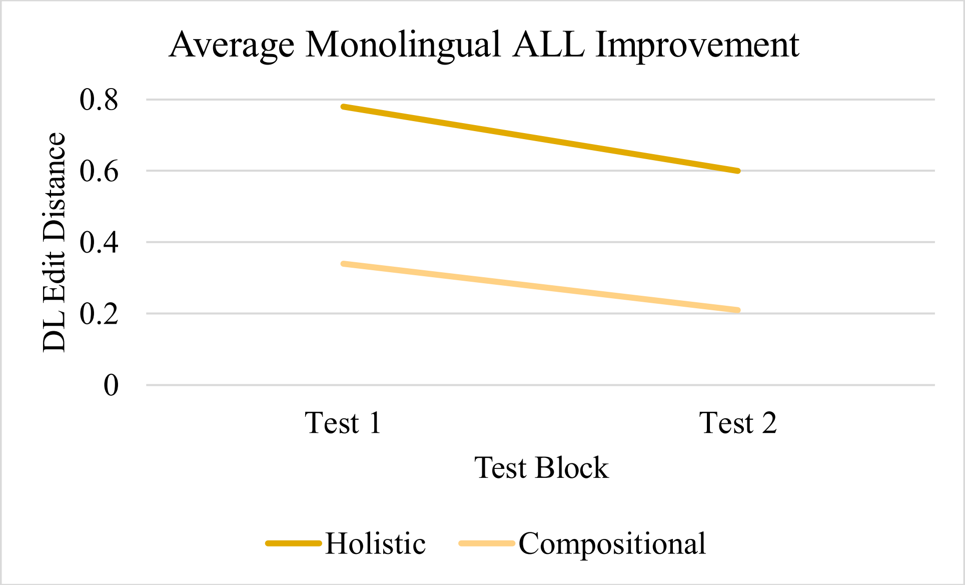


Figure 13 A line graph showing the average monolingual improvement on the ALL task, split by language learned.

This trend continues with the L1 English bilingual group. Similarly to the monolinguals, the average edit distance score for the compositional group was exactly half that of the holistic group for the first test, again decreasing further to under a third of the holistic score by the second task.

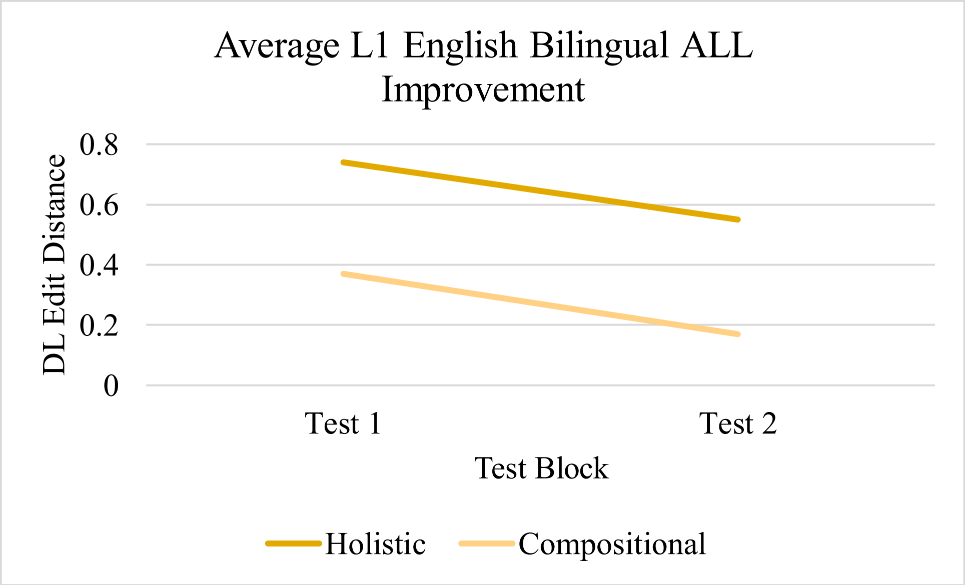


Figure 14 A line graph showing the average L1 English bilingual improvement on the ALL task, split by language learned.

This pattern remains consistent across all three populations. The L2 English bilinguals in the compositional group had an edit distance slightly over half that of their holistic counterpart, once more decreasing further to under a third of the holistic score by the second task. These results and the cause behind them are analysed further throughout the next section.

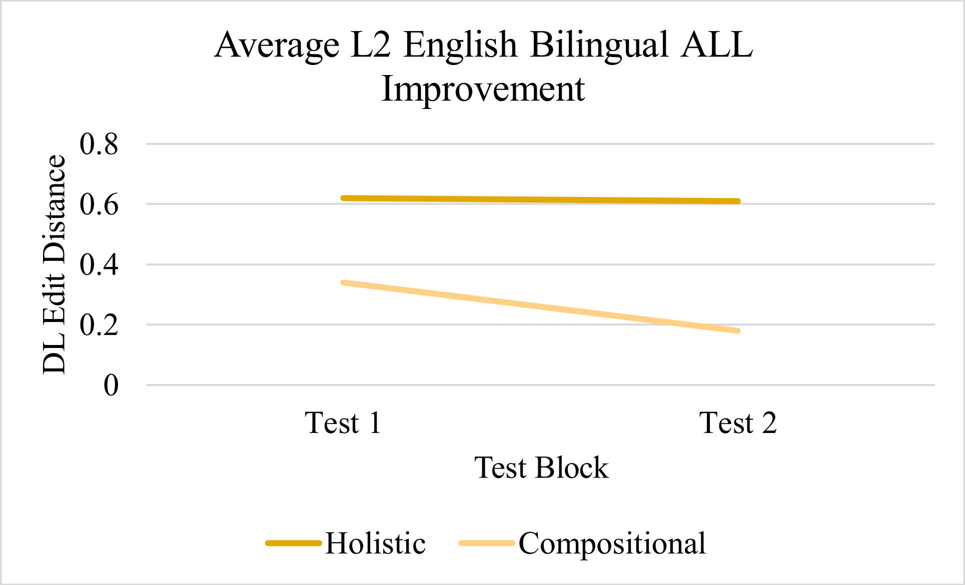


Figure 15 A line graph showing the average L2 English bilingual improvement on the ALL task, split by language learned.

5. Discussion

5.1 The Stroop Tasks

Before anything else, it is important to address the overwhelming amount of anomalous Stroop data. Of the 50 participants, 29 recorded a negative Stroop effect on the first task, this number rising to 31 of the participants by the second Stroop task with 9 of the 29 participants that initially recorded a negative Stroop effect showing a standard Stroop effect on the second task. Haaf and Rouder (2019: 773) point out 'the impossibility of negative Stroop effects', indicating that the way in which the Stroop tasks were constructed in this study was somehow unsuccessful. MacLeod (1991: 167) refers to the version of Stroop task used in this study as a 'colour-word sorting task'. Given that the task was not traditional in the sense that participants did not have to physically say the name of the ink colour, it was known that the RT would be less than Stroop tasks where they did have to do this. However, many Stroop tasks have been carried out without a spoken response and this was deemed acceptable as it would allow the experiment to be carried out by more people, as they could perform it in their own home on their own personal computers. The crux of the issue with this Stroop task is the *type* of possible response. In the other models mentioned in this essay, such as PsyToolkit (2021), participants still had to summon the ink colour name as their response, responding using the corresponding key for the first letter of the ink colour name. This means that participants still had to fully inhibit the word in front of them so as to facilitate the appropriate colour name. In the version carried out in this study, the participants only had to make a judgement on whether the ink colour and word matched or not, and it seems that this method has managed to side-step the Stroop effect in a way. Participants did not have to inhibit the word written before them at all, rather they could focus on the written word and then make a judgement on whether that word matched the ink colour. They did not have to facilitate the ink colour name at all in incongruous cases, as long as they identified that the written word did not match the colour in which it was written. It becomes clear why this is termed a colour-word sorting task now, in that participants are not following the normal Stroop processes but instead sorting stimuli. Participants could simply see stimuli as a 'correct' or 'incorrect'/ 'match' or 'no match' decision and press the appropriate button accordingly, without ever consciously thinking of the name of the ink unless it matched the printed word. This realisation goes some way to explaining the unpredicted data. Were the opportunity to arise, it would be interesting to carry out this experiment again with a more appropriate Stroop task design to truly test the hypotheses.

The unpredicted results of decreasing accuracy in the Stroop task by the compositional group may be caused by some kind of fatigue effect. It is not surprising that this experiment became taxing for participants. While creating an experiment that rendered participants unable to perform with consistent effort throughout is entirely counterproductive, the experiment needed to be as long and as difficult as it was in order to collect the relevant data. Any less time or any easier task would not have targeted the processes that needed to be targeted, or would not have produced enough data to make up a full study. Potentially, this fatigue effect could have been avoided if there were better compensation for carrying out the experiment, such as paying participants for their time. Under those conditions, participants might feel more enthused or dedicated to the experiment. However, that was not an option for this study.

As expected, shown in the beginning of the ALL task results section, on average, participants improved regardless of whether they were operating under the holistic or compositional condition. This, at least, shows that the ALL task was effective for its purpose. In line with the hypothesis that the compositional language would be more learnable, by the second test, the compositional group participant with the highest average edit distance score (0.45) still had a lower score than the average holistic group edit distance (0.58). This supports Kirby et al.'s (2008, 2015) statement that structured languages are more learnable than holistic languages.

Of all the groups, the L2 English bilinguals did not perform as well as expected on the ALL or Stroop tasks given that they were hypothesised to outperform all other groups. However, there were only 7 L2 English bilinguals, 3 of whom learnt the holistic language while the other 4 learnt the compositional. Due to the smaller group size in comparison with the others, one bad score can have a much bigger effect on a group of this size. Across all groups, there were instances of individuals scoring worse than average which is why three different sets of values were given throughout the results. By giving the mean, range and median values, the aim is, in conjunction with the visual representations of the results, to show the data in the most explicit way. Short of including every single piece of raw data in this essay, it is felt that the most responsible and honest way to showcase the results is to give as much relevant information about them as possible. Given that, in some cases, the averages were skewed by a few participants having higher/lower results, the other measures are employed to show this. Another option would be to exclude results that sit a lot higher/lower than the average, but these results did not feel so high or so low as to seem totally anomalous and they still make up important data. Were this study to be carried out again, it might produce more reliable results to control the group sizes more, for example having an equal number of L2 English speakers to the L1 English bilingual group. It would also be worth finding a measure to quantify the bilingualism of the L1 English 'bilingual' group other than their own self-assessment.

6. Conclusion

In this piece of work, a group of 50 participants were trained on two artificial languages, each participant either learning a more holistic or more compositional language. Stroop tasks were employed to measure participants' inhibitory control before and after the ALL task under the hypothesis that the novel linguistic mappings would temporarily increase participants' inhibitory control, manifesting as a decreased Stroop effect in comparison with the one measured in the first task. It was expected that the group trained on the compositional language would have an even further decreased Stroop effect than the holistic group following literature indicating that compositional languages were easier to learn. Whichever participants learned more effectively, the stronger their novel mappings would be, leading to temporarily heightened inhibitory control. However, due to issues with the design of the Stroop task, the hypothesised decreased Stroop effect could not be observed, and the data was inconclusive. The hypothesis around the learnability of compositional languages in comparison with holistic languages was supported by evidence from the ALL task. Possible amendments to the method have been suggested were this research to be redone, with the outlook that more salient results could be gathered with these improvements.

References

Kirby, S., Tamariz, M., Cornish, H. and Smith, K. (2015) ‘Compression and communication in the cultural evolution of linguistic structure’, *Cognition*, 141: 87 – 102.

Dalrymple-Alford, E. C. and Budayr, B. (1966) ‘Examination of some aspects of the Stroop color-word test’, *Perceptual and Motor Skills*, 23: 1211 – 1214.

MacLeod, C. M. (1991) ‘Half a century of research on the Stroop effect: an integrative review’, *Psychological Bulletin*, 109(2): 163 – 203.

Stroop, J. R. (1935) ‘Studies of interference in serial verbal reactions’, *Journal of Experimental Psychology*, 18(6): 643 – 662.

Luk, G., De Sa, E. and Bialystok, E. (2011) ‘Is there a relation between onset age of bilingualism and enhancement of cognitive control’, *Bilingualism: Language and Cognition*, 14: 588–595.

Hermans, D., Bongaerts, T., De Bot, K. and Schreuder, R. (1998) ‘Producing words in a foreign language: can speakers prevent interference from their first language?’ *Bilingualism: Language and Cognition*, 1(3): 213 – 229.

Bialystok, E. (2007) ‘Cognitive effects of bilingualism: how linguistic experience leads to cognitive change’, *International Journal of Bilingual Education and Bilingualism*, 10(3): 210 – 223.

Dick, A. S., Garcia, N. L., Pruden, S. M., Thompson, W. K., Hawes, S. W., Sutherland, M. T., Riedel, M. C., Laird, A. R. and Gonzalez, R. (2019) ‘No evidence for a bilingual executive function advantage in the ABCD study’, *Nature Human Behaviour*, 3: 692 – 701.

Morton, J. B. (2015) ‘Still waiting for real answers’, *Cortex*, 73: 352 – 353.

Kirby, S., Cornish, H. and Smith, K. (2008) ‘Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language’. *Proceedings of the National Academy of Sciences* 105(31): 10681 – 10686.

Barac, R. and Bialystok, E. (2011) ‘Cognitive development of bilingual children’. *Language Teaching* 44(1): 36 – 54.

Peal, E. and Lambert, W. E. (1962) ‘The relation of bilingualism to intelligence’. *Psychological Monographs: General and Applied* 76(27): 1 – 23.

Costa, A., Hernández, M. and Sebastián-Gallés, N. (2008) ‘Bilingualism aids conflict resolution: evidence from the ANT task’. *Cognition* 106(1): 59 – 86.

Czapka, S., Wotschack, C., Klassert, A. and Festman, J. (2020) ‘A path to the bilingual advantage: pairwise matching of individuals’. *Bilingualism: Language and Cognition* 23: 344 – 354.

Reiter, A., Tucha, O. and Lange, K. W. (2005) ‘Executive functions in children with dyslexia’. *Dyslexia* 11(2): 116 – 131.

Klingberg, T., Forssberg, H. and Westerberg, H. (2002) ‘Training of working memory in children with ADHD’. *Journal of Clinical and Experimental Neuropsychology* 24(6): 781 – 791.

Kóbor, A., Takács, A., Bryce, D., Szcűcs, D., Honbolygó, F., Nagy, P. and Csépe, V. (2015) ‘Children with ADHD show impairments in multiple stages of information processing in a Stroop task: an ERP study’. *Developmental Neuropsychology* 40(6): 329 – 347.

Barkley, R. A., Grodzinsky, G. and DuPaul, G. J. (1992) ‘Frontal lobe functions in attention deficit disorder with and without hyperactivity: a new review and research report’. *Journal of Abnormal Child Psychology* 20(2): 163 – 188.

Knight, C. (2018) ‘What is dyslexia? An exploration of the relationship between teachers’ understandings of dyslexia and their training experiences’. *Dyslexia* 24: 207 – 219.

Robinson, S., Goddard, L., Dritschel, B., Wisley, M. and Howlin, P. (2009) ‘Executive functions in children with autism spectrum disorders’. *Brain and Cognition* 71(3): 362 – 368.

Damerau, F. J. (1964) ‘A technique for computer detection and correction of spelling errors’. *Communications of the ACM* 7(3): 171 – 176.

Levenshtein, V. I. (1966) ‘Binary codes capable of correcting deletions, insertions, and reversals’. *Soviet Physics Doklady* 10 (8): 707 – 710.

Zhao, C. and Sahni, S. (2019) ‘String correction using the Damerau-Levenshtein distance’. *7th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS),* Orlando, FL, USA. 19 – 46.

Haaf, J. M. and Rouder, J. N. (2019) ‘Some do and some don’t? Accounting for variability of individual difference structures’, *Pyschonomic Bulletin and Review* 26: 772 – 789.

Cenoz, J. (2013) ‘The influence of bilingualism on third language acquisition: Focus on multilingualism’, *Language Teaching* 46: 71 – 86.

Pear, T. H. (1934) ‘Are linguistic texts accurate?’ *British Journal of Psychology* 25(1): 55 – 62.

Goldstein, S., Naglieri, J. A., Princiotta, D. and Otero, T. M. (2014) ‘Introduction: a history of executive functioning as a theoretical and clinical construct’, in Goldstein, S. and Naglieri, J. A. (eds.) *Handbook of Executive Functioning*. New York: Springer, 3 – 12.

Isquith, P.K., Roth, R.M. and Gioia, G.A. (2014) ‘Assessment of executive functioning using tasks of executive control’, in Goldstein S. and Naglieri, J. (eds) *Handbook of Executive Functioning*. New York: Springer, 333 – 357.

OpenGenus. (2019) ‘OpenGenus IQ: learn computer science – Damerau Levenshtein distance’. Last accessed 29/04/2021, from: https://iq.opengenus.org/damerau-levenshtein-distance/.

PsyToolkit. (2021) ‘Stroop task’. Last accessed 16/02/2021, from: https://www.psytoolkit.org/experiment-library/stroop.html

Stonewall. (2020) ‘Glossary of terms’. Last accessed 05/04/2021, from: https://www.stonewall.org.uk/help-advice/faqs-and-glossary/glossary-terms.

Appendix

Information Sheet

This is the information sheet that participants were presented with at the beginning of the experiment:

Information Sheet

English Literature, Language and Linguistics

Study title:	Artificial language learning and inhibitory control
Principal Investigator:	Dr. Christine Cuskley and Dr. Joel Wallenberg
Researcher collecting current data:	Fionnuala Lynch

What is this document? This document explains what kind of study we're doing, what your rights are, and what will be done with your data. If there are any special benefits or risks, they will be explained here. Please read the information carefully and retain it for your records.

Nature of the study: You are about participate in a study which involves three tasks: a task where you will be shown the name of a colour and asked whether the colour the word is written in matches the name itself; a task in which you are asked to learn a miniature alien language; and finally, the first task again. Your session should last for about 20 – 30 minutes. You will be given full instructions before the study begins.

Compensation: There are no known risks to participation in this study. The only benefits to you personally are those you draw from making a contribution to our knowledge about language.

Confidentiality: The data we collect will not be associated with your name or with any other personal details or identifying information.

Voluntary participation: Your participation is voluntary, and you may stop playing at any time for any reason. Any data you provide or produce up to this point will not be collected or stored. To withdraw after you have completed the study, contact one of the researchers via the emails provided, quoting the number you are given upon completing the study.

Contact information: This research is being conducted by Fionnuala Lynch at Newcastle University and is overseen by Dr. Christine Cuskley and Dr. Joel Wallenberg. The researchers can be contacted at f.lynch@newcastle.ac.uk, christine.cuskley@newcastle.ac.uk or joel.wallenberg@newcastle.ac.uk for questions or to report a research-related problem. Contact Newcastle University Research Ethics at res.policy@ncl.ac.uk if you have concerns regarding your rights as a participant in the research

By agreeing to these terms, you consent:

- **that the anonymous response data you produce may be kept permanently in research archives at Newcastle University, and used for the specific research project which made them.**
- **to your anonymous data being used by the above-named researcher as well as by other qualified researchers, for teaching or research purposes, in professional presentations and publications.**
- **To your anonymous data being included in aggregate data released as part of scholarly publication.**

You have the right to terminate my participation at any point. If you choose to withdraw formally, your data will be deleted.

Debrief

This is the debrief that participants were presented with on completing the experiment:

Debrief

What was the study about?

This study is about investigating what can help or hinder executive function. Executive function is the set of mental skills that assists us with things like concentration and filtering information. It's theorised that this set of skills can be strengthened by language learning (in this task, the little alien language you learned). The task with the coloured words that you did before and after the language task is a way to measure executive function is. We want to measure if people perform better on the second colour-word task than the first one, which could show that the language task helped them.

We suspect that a task with an alien language that is more systematic is more likely to strengthen executive function, so we gave some participants rule governed language where part of each word referred to each shape and part to its pattern (e.g., shape-pattern; ege-wawa, mega-wawa, mega-wawuetc.). Other participants got a more holistic language, where entire words referred to entire shapes (e.g., pihino, gakho, nemone). These shapes and words were adapted from an earlier artificial language learning study (https://www.sciencedirect.com/science/article/pii/S0010027715000815).

Where can I find out more?

The study is conducted by researchers in Linguistics at Newcastle University. In case you missed it, the detailed information sheet about the study can be downloaded here. If you have any further questions about the research, please email f.lynch@newcastle.ac.uk or Christine Cuskley.

Stroop Blocks

These are the two pseudorandomised Stroop blocks used in the experiment:

Order 1				**Order 2**			
Trial Number	**Ink Colour**	**Colour Name**	**Stimulus**	**Trial Number**	**Ink Colour**	**Colour Name**	**Stimulus**
1	Red	Blue	blue	1	Blue	Green	green
2	Red	Orange	orange	2	Blue	Red	red
3	Blue	Blue	blue	3	Orange	Green	green
4	Orange	Orange	orange	4	Blue	Green	green
5	Green	Blue	blue	5	Blue	Orange	orange
6	Red	Blue	blue	6	Green	Green	green
7	Orange	Green	green	7	Orange	Blue	blue
8	Red	Red	red	8	Red	Green	green
9	Green	Blue	blue	9	Orange	Green	green
10	Green	Red	red	10	Blue	Red	red
11	Blue	Blue	blue	11	Orange	Orange	orange
12	Green	Orange	orange	12	Blue	Red	red
13	Red	Blue	blue	13	Blue	Green	green
14	Orange	Orange	orange	14	Blue	Orange	orange
15	Red	Red	red	15	Red	Green	green
16	Orange	Orange	orange	16	Orange	Orange	orange
17	Red	Blue	blue	17	Green	Green	green
18	Green	Orange	orange	18	Orange	Orange	orange

19	Green	Blue	blue	19	Red	Green	green
20	Orange	Orange	orange	20	Red	Blue	blue
21	Green	Green	green	21	Green	Orange	orange
22	Orange	Blue	blue	22	Red	Orange	orange
23	Orange	Green	green	23	Red	Blue	blue
24	Red	Blue	blue	24	Green	Green	green
25	Orange	Blue	blue	25	Orange	Red	red
26	Green	Blue	blue	26	Blue	Blue	blue
27	Green	Red	red	27	Green	Orange	orange
28	Blue	Orange	orange	28	Blue	Blue	blue
29	Red	Orange	orange	29	Orange	Red	red
30	Blue	Blue	blue	30	Blue	Red	red
31	Green	Orange	orange	31	Orange	Green	green
32	Red	Red	red	32	Red	Blue	blue
33	Blue	Orange	orange	33	Red	Orange	orange
34	Green	Green	green	34	Green	Orange	orange
35	Orange	Red	red	35	Red	Red	red
36	Blue	Red	red	36	Blue	Green	green
37	Blue	Orange	orange	37	Red	Green	green
38	Red	Green	green	38	Orange	Blue	blue
39	Red	Orange	orange	39	Red	Red	red
40	Blue	Orange	orange	40	Orange	Green	green
41	Blue	Green	green	41	Orange	Red	red
42	Orange	Green	green	42	Blue	Green	green
43	Blue	Red	red	43	Orange	Green	green
44	Green	Green	green	44	Red	Red	red
45	Red	Orange	orange	45	Green	Blue	blue
46	Red	Green	green	46	Red	Orange	orange
47	Red	Orange	orange	47	Blue	Blue	blue
48	Blue	Blue	blue	48	Green	Orange	orange
49	Green	Red	red	49	Blue	Red	red
50	Green	Orange	orange	50	Orange	Orange	orange
51	Red	Red	red	51	Blue	Blue	blue
52	Blue	Blue	blue	52	Red	Green	green
53	Orange	Red	red	53	Orange	Blue	blue
54	Blue	Red	red	54	Red	Red	red
55	Green	Green	green	55	Blue	Orange	orange
56	Orange	Red	red	56	Green	Green	green
57	Orange	Green	green	57	Blue	Orange	orange
58	Blue	Red	red	58	Green	Green	green
59	Blue	Green	green	59	Orange	Blue	blue
60	Blue	Orange	orange	60	Orange	Red	red
61	Blue	Green	green	61	Green	Blue	blue
62	Red	Green	green	62	Green	Red	red
63	Orange	Blue	blue	63	Blue	Orange	orange
64	Green	Green	green	64	Red	Red	red
65	Orange	Red	red	65	Blue	Blue	blue
66	Blue	Green	green	66	Red	Orange	orange
67	Red	Green	green	67	Green	Orange	orange
68	Orange	Blue	blue	68	Green	Blue	blue
69	Green	Red	red	69	Red	Blue	blue
70	Green	Orange	orange	70	Orange	Blue	blue
71	Green	Blue	blue	71	Green	Red	red

72	Green	Red	red	72	Green	Blue	blue
73	Orange	Orange	orange	73	Red	Orange	orange
74	Blue	Green	green	74	Red	Blue	blue
75	Orange	Red	red	75	Green	Blue	blue
76	Blue	Red	red	76	Green	Red	red
77	Orange	Green	green	77	Orange	Red	red
78	Red	Green	green	78	Green	Red	red
79	Orange	Blue	blue	79	Orange	Orange	orange
80	Red	Red	red	80	Green	Red	red