Universidade do Minho
Escola de Psicologia

Luís Carlos Morais Teixeira

Is the Foreign Language effect modulated by emotionality? A study using a base-rate task.

abril de 2023

Is the Foreign Language effect modulated by emotionality? A study using a base-rate task.
Luís Teixeira
UMinho | 2023

Universidade do Minho
Escola de Psicologia

Luís Carlos Morais Teixeira

Is the Foreign Language effect modulated by emotionality? A study using a base-rate task.

Dissertação de Mestrado
Mestrado Integrdo em Psicologia

Trabalho efetuado sob a orientação da
Professora Doutora Montserrat Comesaña
e do
Professor Doutor Armando Machado

abril de 2023

Agradecimentos

Em primeiro lugar quero agradecer à minha família sem a qual tirar um Mestrado teria sido impossível. À minha mãe por nunca deixar de acreditar em mim, ao meu irmão Luís pelos sacrifícios que fez para que eu pudesse terminar o curso, ao meu irmão David por todas as vezes em que reviu os meus textos ou perdeu tempo a ouvir-me divagar sobre matéria, e por último ao meu pai por estar presente à maneira dele.

Gostaria ainda de agradecer a professora Doutora Montserrat Comesaña e ao professor Doutor Armando Machado pelo seu trabalho como orientadores, o seu feedback foi uma ferramenta indispensável ao longo de todo o processo e acredito que fez de mim um profissional mais completo e competente. Queria ainda agradecer ao professor Juan Haro pela sua colaboração neste projeto.

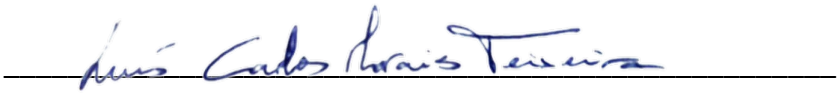
E por último, gostaria de agradecer ao staff da MultiCris pelo serviço impecável que me tem prestado ao longo de toda esta jornada.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Braga, 26 of April 2023.



(Luís Carlos Morais Teixeira)

O efeito da Língua Estrangeira é modulado pela emoção? Um estudo usando uma tarefa de taxa base.

Resumo

O efeito da Língua Estrangeira (FLe) é o fenómeno pelo qual pensar numa língua não nativa afeta os processos de tomada de decisão. O FLe tem sido extensivamente estudado nos domínios de perda e aversão ao risco, bem como na tomada de decisão moral, mostrando que o uso de língua estrangeira geralmente reduz a aversão ao risco no primeiro e torna os sujeitos mais dispostos a aceitar danos para maximizar os resultados no segundo. Isso levou os investigadores a acreditar que pensar numa língua estrangeira pode reduzir os vieses de decisão muito provavelmente porque envolve um processamento menos automático e mais controlado, especialmente quando a tarefa mobiliza o sistema afetivo. No entanto, as evidências nem sempre são consistentes com essa visão mostrando justamente o efeito contrário, por exemplo, pior desempenho no raciocínio lógico formal ou ausência de diferenças a nível de desempenho executivo em crianças. No presente estudo, utilizamos uma tarefa de taxa de base modificada para testar o FLe em bilíngues luso-europeus de nível intermédio a alto de proficiência em inglês. Nesta tarefa, os participantes são informados sobre uma taxa de base para pertencer a um determinado grupo e informações heurísticas e estereotipadas sobre um indivíduo específico. Em seguida, devem avaliar a probabilidade de que esse indivíduo pertença a um dos grupos, ignorando as informações salientes, intuitivas e estereotipadas também fornecidas. O conteúdo afetivo das informações heurísticas e o consequente grau de conflito entre intuições concorrentes (baixo vs. alto), bem como a língua em que a tarefa foi apresentada (nativa vs. estrangeira), foram manipulados. Os resultados não revelaram efeito significativo da língua nas estimativas de probabilidade ($p > .05$). Estes dados sugerem que o uso de uma língua estrangeira pode não ter um impacto significativo na redução do viés de representatividade na nossa população. Também medimos a confiança dos participantes nas suas respostas e, novamente, não encontramos um FLe. Em vez disso, a confiança parece depender mais da idade, da idade de aquisição da segunda língua (AoA) e da proficiência na língua estrangeira dos participantes.

Palavras-chave: bilinguismo, efeito da língua estrangeira, taxa base, emoção, deteção de conflito.

Is the Foreign Language effect modulated by emotionality? A study using a base-rate task.

Abstract

The Foreign Language effect (FLe) is the phenomenon by which thinking in a non-native language affects decision-making processes. The FLe has been extensively studied in the domains of loss and risk aversion, as well as in moral decision-making, showing that the use of a foreign language generally reduces risk aversion in the former and makes subjects more willing to accept harm to maximize outcomes in the latter. This has led researchers to believe that thinking in a foreign language can reduce decision-making biases, most likely because it involves less automatic and more controlled processing, especially when the task mobilizes the affective system. However, the evidence is not always consistent with this view, showing precisely the opposite effect, i.e., worse performance in formal logical reasoning, absence of differences in executive performance in children. In the present study, we used a modified base-rate task to test the FLe in European Portuguese-English bilinguals with intermediate to high level of English proficiency. In this task participants are given a base-rate for belonging to a certain group and heuristic and stereotyped information about a certain individual. Then, they must evaluate the probability that that person belongs to one of the groups, whilst ignoring the salient, intuitive, and stereotyped information also provided. The affective content of the heuristic information and consequent degree of conflict between competing intuitions (low vs.high) as well as the language in which the task was presented (native vs. foreign) were manipulated. The results revealed no significant effect of language on probability estimates ($p > .05$). These findings suggest that foreign language usage may not have a significant impact on reducing the representativeness bias in this population. We also measured participants confidence in their answers and again, we did not find a FLe. Instead, confidence appears to be more dependent on the age, age of acquisition of second language (AoA) and foreign language proficiency of the participants.

Keywords: bilingualism, foreign language effect, base-rate, emotion, conflict detection.

Table of Contents

List of Figures

List of Tables

List of acronyms

NL – Native Language
FL – Foreign Language
LC – Low Conflict
HC – High Conflict
AoA – Age of Acquisition of FL
SIM – Stunted Intuitions Model
CTM – Conflict Threshold Model

Introduction

The Foreign language effect (FLe), also referred to as bilingual advantage or bilingual effect, is the modulation of the cognitive processes underlying judgment and decision-making when thinking in a non-native language (Keysar et al., 2012). The FLe has made the rounds in the field of cognitive psychology and neuropsychology. Inspired by a Dual-Process approach to cognition (Kahneman, 2003), the original hypothesis was that thinking in a foreign language affects people's decision-making processes by reducing biases. For instance, risk aversion for gains and risk seeking for losses, arguably the result of a framing effect, seems to be reduced when choices are presented in the second or foreign language. The reduction occurs probably because there is a more controlled processing of this language when compared with the more intuitive processing engaged in the native language (see Keysar et al., 2012; also, Costa et al., 2014).

The Dual-Process Theory framework assumes that cognition consists of two complementary processes: Type I processes which are fast, autonomous, and do not require access to working memory; and Type II processes which require access to working memory to deliberate and produce rational responses, and therefore are also more effortful. The most prominent model of Dual-Process Theory proposes that people tend to use Type I processes by default, and engage in Type II reasoning only when they detect a conflict between two or more intuitive responses, or if no intuitive response is readily available (Białek et al., 2020; Kahneman, 2012; Keysar et al., 2012). Considering this framework two general predictions regarding the FLe were made. On the one hand, given that speakers are usually less fluent in a foreign language, using it signals that more deliberative processes are needed – making people rely even more on systematic processes and thereby reduce decision biases. On the other hand, the FLe could have the opposite effect since processing in a foreign language might be more cognitively demanding or costly. The increase in cognitive load might lead to greater reliance on intuitive and affective processes. If such a reduced systematicity account were true, then the use of a foreign language should exacerbate certain decision biases that arise from heuristics and affective/emotional processes.

These hypotheses were initially tested by Keysar et al. (2012) in the field of loss and risk-aversion using framing (e.g., Asian Disease problem) and gambling experiences. In one of the gambling experiences, loss aversion was measured by the number of bets a participant was willing to make whilst the value of the bet and/or its odds of winning varied. The results showed that overall participants who received the instructions in their foreign language were more likely to accept the bet than those who received them in their native language (NL), which implies that the former were either less sensitive to the possibility of a loss or less sensitive to the loss itself.

In the Asian disease problem, participants were presented with a hypothetical scenario: "Recently, a dangerous new disease has been going around. Without medicine, 600,000 people will die from it. In order to save these people, two types of medicine are being made" and asked to make a binary choice between a *safe* and a *risky* option with the same expected value. Critically, the same scenario could be framed as a perceived Gain[1] or Loss[2], and the framing influenced the choices people made – participants tended to choose the *safe* option more often when the scenario was framed as a Gain and that preference shifted when the problem was framed as a Loss. This frame-dependent choice is a classic example of people's tendency to avoid perceived losses, also known as loss-aversion (Kahneman & Tversky, 1979; Kahneman, 2012). However, when performing in a foreign language, participants were no longer sensitive to the framing condition, suggesting that "the robust asymmetry in risk preferences disappears when a decision takes place in a foreign language" (Keysar et al., 2012; replicated in Costa, et al., 2014a). Overall, these results led the authors to propose that people rely more on systematic processes that respect normative rules when making decisions in a foreign language than when making decisions in their native tongue.

The FLe was also extensively studied in the domain of moral decisions. Many studies have addressed how being in a foreign language context can affect decision-making when there is the possibility to accept harm to maximize outcomes (Cipolletti et al., 2016; Costa et al., 2014b; Dylman & Champoux-Larsson, 2020; Miozzo et al., 2020). As a piece of evidence, Costa et al. (2014b) showed that participants presented with the "footbridge" trolley dilemma, in which participants choose between doing nothing and letting 5 people die, or pushing a person to the track, sacrificing one life to save five others, are more willing to make the sacrifice when the dilemma is presented in a foreign language than in their native language. The authors explain their results by saying that: "The reduction of emotionality elicited by a foreign language may promote psychological distance in general. Increasing psychological distance leads individuals to construe situations in more abstract terms, which in some situations aligns with more utilitarian decision making." (Costa et al., 2014b, p.5).

Taken together, the research in moral decision-making and risk-aversion domains has shown that foreign language usage can affect participant's decisions compared to native language, making them more willing to accept harm to maximize outcomes in the former and reducing risk aversion in the latter.

[1] If you choose Medicine A, 200,000 people will be saved. If you choose Medicine B, there is a 33.3% chance that 600,000 people will be saved and a 66.6% chance that no one will be saved".

[2] If you choose Medicine A, 400,000 people will die. If you choose Medicine B, there is a 33.3% chance that no one will die and a 66.6% chance that 600,000 people will die."

As mentioned above, these findings sustain the view that people rely more on systematic (rule-based) processes, characteristic of the system II, when using a foreign language.

However, the evidence is not always consistent with this view. For instance, Bialek et al. (2020) showed that participants were less accurate at identifying invalid (classic/Socratic) syllogisms when presented in a foreign language. Thus, contrary to what was observed in the domain of moral decisions and of loss and risk-aversion using framing and gambling experiences, the results here observed suggest that using a foreign language may impair the participant's ability to detect the conflict between competing intuitions and allocate cognitive effort, therefore preventing them from engaging in the type II processes that would be necessary/required to provide the correct answer. These findings led the authors to propose a couple of new models (i.e., The Conflict Threshold Model (CTM) and the Stunted Intuitions Model (SIM)) based on conflict detection, both incompatible with the cognition-based account mentioned above. As the authors put it:

> "It is difficult to imagine reasoning in a foreign language being both: 1) more influenced by Type II reflective processes, and 2) less accurate in syllogistic reasoning and less amenable to conflict detection." (Białek et al., 2020, p.12).

The models are based on the idea that that the difference in accuracy between native and foreign language (FL) reasoners occurs before engaging in Type II processes. The Conflict Threshold Model essentially suggests that FL reasoners are less sensitive to conflict and require a higher degree of it before they can detect it and allocate cognitive effort accordingly. An alternative explanation is the Stunted Intuitions Model, according to which, reasoning in a foreign language can inhibit or "stunt" one's logical intuitions, either to the point where they are non-existent or simply overwhelmed by heuristic-based intuitions. This model predicts that FL reasoners' intuitions might conflict less with one another, which could lead to decreased reasoning accuracy compared to native language (NL) reasoners.

Another concurrent explanation for the FLe proposes that it stems from decreased emotional resonance. If so, the results obtained in moral decision-making and risk-aversion would not be due to enhanced systematicity but to decreased susceptibility to biases introduced by emotional processing, which may create the appearance of increased systematicity. Bialek et al. (2020) claim that their Conflict Threshold Model:

> "(I)s not necessarily irreconcilable with the emotion-based mechanism (Costa, Foucart, Hayakawa, et al., 2014; Keysar et al., 2012; Vives et al., 2018). For example, De Neys Moyens, and Ansteenwegen (2010) proposed that conflict detection is affective by nature. If it is the case that foreign language reasoners experience diminished emotion as a product of reasoning in their

foreign language (Pavlenko, 2008, 2012), and the experience of emotion is crucial to the detection of conflict, it follows that they will be less able to detect conflict even when it is present." (p.12).

Further evidence for this explanation can be found in another experiment by Costa et al. (2014b) using the "switch" dilemma, which is similar to the "footbridge" dilemma in all respects but one: Participants are not asked to push the man into the track, but instead to pull a switch that will divert the train killing one man that is on the tracks. The results showed not only an increase in the number of utilitarian choices (i.e., pull a switch, killing the man on the tracks) in both the NL and FL participants but also that the asymmetry between these two groups was no longer present. The authors explain these results by saying that "People are more willing to sacrifice the one man by pulling the switch than by pushing him off the footbridge, and one of the primary reasons is that pulling the switch is less emotionally aversive (Greene et al., 2001)." (Costa et al., 2014b, p.3). The reduced emotionality explanation also provides the additional benefit of helping to account for the cases when no effect was found, which generally are purely cognitive tasks like the Cognitive Reflection Test (CRT) (Frederick, 2005). Given that the task is strictly cognitive, the reduced emotionality afforded by the FL accrues no benefit, and, in some cases, the reduced fluency can even be detrimental like in syllogistic reasoning.

In addition to the foregoing experiments, the reduced emotional response explanation is also supported by some studies in which participants rated words as less emotional in the second than the first language (especially negative and taboo words), and by studies that measured skin conductance responses and event-related potentials, which also found a diminished emotional response when using the second language (Harris et al., 2003; see also Rosselli et al., 2017 for an overview).

Another source of evidence against the idea that thinking in a foreign language activates Type 2 systematic processing can be found in Lowe et al. (2021) meta-analytic review, which found no FLe when performing cognitive related tasks. This review was done with children though, and thus differences can be attributed to age-related effects.

It's also important to note that the FLe seems to depend on other factors including cultural influences and the relative distance between the two languages that bilinguals speak (linguistically closer vs. distant). Dylman and Champoux-Larsson (2020) found a FLe in a framing task (i.e., the Asian Disease problem) and in the "footbridge" dilemma when the foreign language had low cultural influence (French for Swedish-French bilinguals), but not when it had high cultural influence (English for Swedish-English

bilinguals).[3] With respect to language distance, Miozzo et al. (2020) compared the use of Italian with other Italian local dialects, and found that the FLe is present when the languages are different even though they are both used in the same country. On the other side of the spectrum, Dylman and Champoux-Larsson (2020) report no Fle in linguistically similar languages (Swedish – Norwegian)[4]. Therefore, more research is clearly needed to get a clearer picture of the robustness of the FLe and the mechanisms underlying it.

A suitable task to put into test the reduced emotionality account is a modified base-rate task inspired in Kahneman and Tversly (1973). In a typical base-rate task, participants must assess the probability of a person belonging to a particular group while ignoring the salient, intuitive, stereotypical information they have about the person. In Kahneman and Tversky's original study, base rate probabilities were stated in advance: Participants were informed that in a total of 100 individuals, 30 were engineers and 70 were lawyers. Next, they read a description of the personality of an individual randomly drawn from this population (e.g., "Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing and mathematical puzzles."). Finally, they had to estimate the probability that the individual described was an engineer.

Kahneman and Tversky (1973) believed that contrary to the common-sense view of human thinking, our predictions are not always rooted in rational and deliberate reasoning. Instead, they argued, people are easily influenced by non-relevant information simply because it resembles a mental category or stereotype – *representativeness heuristic* – while ignoring all other pertinent information such as base-rate probabilities and the validity of additional information. If probability estimations were purely logic and rational, one would expect that regardless of the personality description, the participants would respond that the probability is 30%. However, if the representativeness heuristic is engaged, this value would be expected to vary as a function of the stereotype activated by the characteristics included in the personality description (which in the example above is the stereotype of an engineer). Kahneman and Tversky found that people do indeed estimate probabilities based on the description rather than on base rates.

Critically the conflict level between the base-rate and the description can be manipulated. For example, we might vary the base-rates making them congruent or incongruent with the description. Vives

[3] Although, no precise definition of *cultural influence* was given, it was implied that social and other types of media may be variables to consider.

[4] No objective measures of language proximity were used. The authors literally state that: "the linguistic similarity between Swedish and Norwegian means that the 2 languages can be understood to a certain degree across borders, and Swedes and Norwegians speaking to each other using each of their native languages can understand one another to a certain extent." (Dylman & Champoux-Larsson, 2020, p. 5)

et al. (2018) used three kinds of stimuli that differed in the relationship between the base-rate information and the description: a) one was Congruent (the base-rate and the description both pointed to the same response), b) one was Neutral (the description did not point to any response), and c) one was Incongruent (the base-rate pointed to one response and the description to another). This feature allowed the authors to see whether people weighted differently the base-rate and stereotypical information when using their native or foreign language. Compared to the neutral condition, participants showed significant higher accuracy in the congruent item and significant lower accuracy in the incongruent item, with a significant effect size. The use of the representativeness heuristic was not influenced by language context, as indicated by the absence of the main effect or any interaction with this factor.

However, Vives et al. (2018) asked their participants a dichotomous question i.e., is the person you read about: a) an engineer or b) a lawyer. This meant that the focus of their study was the accuracy of the participants. We took a slightly different approach; looking at the differences in accuracy reported by Vives et al. (2018) and the classic results from Kahneman (2012); Kahneman & Tversky (1973, 1979), we know subjects tend to be more inaccurate when the description they are presented with is incongruent with the base-rate, but how much? Thus, we were interested in measuring the size of the inaccuracy.

Knowing if there are differences in the size of the overestimations made between NL and FL reasoners might provide further evidence for one of the models proposed by Białek et al. (2020) and further understand the role of emotions in the FLe. Given that we focused on incongruent stimuli, we kept the base-rate 30:70 constant across all experimental conditions and two descriptions that were incongruent with the base-rate were used. Critically, they varied only in the level of emotional activation (considering the valence and arousal of the words used within the description). This way we hoped to control the level of conflict between intuitions across conditions by varying the emotional characteristics of the stimuli (i.e., description). Following the reasoning of Białek et al. (2020) we created a setup which consisted of a low conflict condition (i.e., base-rate 30:70 paired with incongruent description with low emotional activation) and of a high conflict condition (i.e., base-rate 30:70 paired with incongruent description with high emotional activation). Our stimuli were design so that the bigger the emotional activation elicited by the stimuli, the more incongruent with the base-rate it should be, and therefore more conflict should be generated. In summary, there are two sources of conflict in our task, one is constant (the base-rate is always incongruent with the description), and the second is variable (emotional description).

Another interesting variable to be measured besides probability estimation is confidence in the answer given. This is because confidence appears to be directly related to congruence and coherence between inputs or stimuli rather than rational processes. As Kahneman and Tversky (1973) put it:

> "The more consistent the input, the more representative the predicted score will appear and the greater the confidence in that prediction. [...] The intuition that consistent profiles allow greater predictability than inconsistent profiles is compelling. It is worth noting, however, that this belief is incompatible with the commonly applied multivariate model of prediction (i.e., the normal linear model) in which predictive accuracy is independent of within-profile variability. [...] Thus, a paradoxical situation arises where high intercorrelations among inputs increase confidence and decrease validity" (p.249).

This is precisely what we have found in a preliminary unpublished study involving a base-rate task, i.e., participants who gave the correct answer reported a lower degree of confidence than the participants who gave the incorrect answer.

In sum, in the present base-rate study we seek to understand the mechanisms underlying the FLe via the manipulation of emotionality of the description (low conflict vs high conflict) and the language in which the description is given (native vs foreign). Three dependent variables were registered, namely, the *estimated probability*, *confidence ratings* and *bet sizes*.

Depending on which model we choose we should expect different results. According to the Stunted Intuitions Model, there would be a FLe for both high and low conflict base-rate problems, since intuitions relying on the base-rate information would be limited in both cases. In our task this implies that participants who perform the task in their NL will perform better than those who do not, meaning that we should see larger probability estimates when the task is performed in the FL compared to the NL regardless of the experimental condition relating to emotion/conflict.

In contrast, the Conflict Threshold Model suggests that when the conflict between base-rate and heuristic information (description) becomes prominent enough (high conflict condition), FL reasoners will be able to recognize it. Therefore, two outcomes could be expected, if we found that NL participants outperform FL participants in the low conflict condition, those differences should disappear in the high conflict condition. Conversely, if NL and FL participants showed no differences in the low conflict condition, it would be expected to see a difference favouring FL participants in the high conflict condition. Lastly, regarding the confidence ratings, based on (Kahneman & Tversky, 1973) and our preliminary unpublished study we expected that participants who gave the correct answer would report lower levels of confidence than those who do not. If the conflict between the intuitions generated by the base-rate and

heuristic information is not detected, meaning the participant provides an answer according to the heuristic information (incorrect answer), confidence should be higher than for those who provide an answer according to the base-rate (correct answer). As previously mentioned, that should happen because the greater the congruency between inputs, the higher the confidence, regardless of accuracy. Since no data is available relating to how confidence changes with the language used (native vs foreign), we can only expect that the confidence ratings will be the highest when the least amount of conflict is detected and the lowest when the most amount of conflict is detected. That is to say that confidence should be lower when participants follow the base-rate then when they follow the heuristic information given (i.e., description).

Methods

Participants

Two hundred and seventeen European Portuguese (NL) – English (FL) bilinguals with intermediate to high proficiency (164_{female}, M_{age} = 22.8; SD =7.9) took part in the experiment voluntarily. Participants were recruited from several Portuguese Universities (University of Minho n=186, other Universities n = 20, and no affiliation n = 11). Students from the University of Minho received extra credits for participating in the study. All of them had normal or corrected-to-normal vision and signed a consent form for the research before filling out a sociolinguistic questionnaire (Li et al., 2020) and an online lexical decision task – LexTALE (Lemhöfer, K., & Broersma, M., 2012) – which allowed us to know the degree of FL proficiency of participants. Data from the questionnaire revealed that the average age of acquisition of English as a FL was 9.79 years (SD = 3.99). The mean of subjective ratings of reading, writing, speaking, and listening skills in the FL, based on a 7-point Likert scale (from 1 = very poor to 7 = native-like), was 5.3 (SD = 1). In the LexTALE task the average score was 72.14 out of 100 (*SD* = 10.71). LexTALE was used since it has been shown that its scores can be good predictors of FL vocabulary knowledge and to give a fair indication of general English proficiency.

Materials

Given Covid-19 related constraints, the experiment was conducted online using the JsPsych software (de Leeuw, 2015). The experimental task materials consisted of an introduction text that was identical for all participants in a given language condition: a) FL: *"A team of statisticians working for a company analysed the injury reports from the previous year. They interviewed 30 workers who had an*

injury and 70 workers who had not. During the interview, the workers were asked if they had any other injuries in the past. Based on their answers, brief reports were written about the 30 workers who had had an injury and the 70 workers who had not had an injury. You will see next one of those reports chosen at random from the 100 available descriptions.", and b) NL: "Uma equipa de estatísticos que trabalham para uma empresa analisou relatórios de lesões do ano anterior. Eles entrevistaram 30 trabalhadores que tiveram uma lesão e 70 trabalhadores que não tiveram. Nessas entrevistas, os trabalhadores foram questionados se tiveram outras lesões no passado. Com base nisso, pequenos relatórios foram escritos sobre os 30 trabalhadores que sofreram lesões e os 70 trabalhadores que não sofreram. A seguir, verá uma parte de um desses relatórios escolhidos aleatoriamente entre os 100 disponíveis.".

Afterwards, a description that was condition specific according to the experimental condition was presented: **a)** low conflict in FL*: "J. Brady is 34 years old. He has been working with the company for 5 years. His record shows he is a good employee; he fulfils his duties in a timely manner. When he was young, he fell from a second floor which resulted in a fracture of his right arm and a cut on his forehead."*; **b)** high conflict in FL*: "J. Brady is 34 years old. He has been working with the company for 5 years. His record shows he is a good employee; he fulfils his duties in a timely manner. When he was young, he fell from a second floor which resulted in a massive, exposed fracture of his right arm and a blood gushing laceration on his forehead.";* **c)** low conflict in NL*: "O Sr. Pereira tem 34 anos. Trabalha na empresa há 5 anos. O seu histórico mostra que ele é um bom funcionário, que cumpre suas funções em tempo útil. Quando era jovem caiu do segundo andar, o que resultou numa fratura no braço direito e um corte na cabeça."* and **d)** high conflict in NL*: "O Sr. Pereira tem 34 anos. Trabalha na empresa há 5 anos. O seu histórico mostra que ele é um bom funcionário, que cumpre suas funções em tempo útil. Quando ele era jovem caiu do segundo andar, o que resultou numa devastadora fratura exposta no braço direito, e ainda numa enorme e sangrenta laceração na cabeça.".*

To ensure that our descriptions were perceived as emotionally different, we conducted a preliminary study where participants were asked to rate on a scale the valence and arousal of the stimuli used in the four experimental conditions. The valence scale ranged from 1 (unhappy) to 9 (happy). Regarding the FL condition, the 18 participants who read the low conflict version of the description compared to the 19 participants who read the high conflict version reported significantly higher valence scores. That is, the stimulus in the low conflict condition was perceived as more positive than the high conflict (mean rating 4.5 vs. 3.1, $t(35) = 2.8$, $p = .004$). In the NL condition, the 20 participants who read the low conflict version of the description compared to the 17 participants who read the high conflict version reported similar valence scores (mean rating 3.85 vs. 3.7, $t(35) = 0.345$, $p = .336$). The arousal

scale also ranged from 1 (calm) to 9 (aroused). In the FL condition, the 18 participants who read the low conflict version of the description compared to the 19 participants who read the high conflict version reported similar scores of arousal (mean rating 4.5 vs. 4.6, t(35) = -0.197, p = .423). In the NL condition, the 20 participants who read the low conflict version of the description compared to the 17 participants who read the high conflict version reported significantly lower scores of arousal, meaning that the stimulus in the low conflict condition was perceived as less arousing than in the high conflict condition (mean rating 3.85 vs. 4.88, t(35) = -2.240, p = .016).

Considering the data from the above-mentioned preliminary study involving a base-rate task, we decided to use two measures of confidence: a) a classic measure of confidence using a Likert scale; and b) an additional measure of confidence, in which participants will be asked to partake in a hypothetical bet. In that preliminary study this measure proved to be more representative than the classic Likert scale. In the classic Likert scale the average *confidence rating* for all the participants was near the middle point of the scale; when betting however, the *bet size* appeared to be dependent on whether the participants got the correct answer to the probability estimation or not (the correct answer being the one that follows the base-rate, i.e., "30%"). The confidence was measured with a Likert scale that ranged from 1 (not confident at all) to 5 (extremely confident); and with the hypothetical bet, also using a Likert scale ranging from 1 to 128€ with geometric increments of ratio 2; the hypothetical winnings being twice the bet. This meant that the bigger the bet, the bigger its expected value. The expected value of the bets ranged from 0.5 to 32.

Procedure

Participants were sent a link that would take them to the informed consent form. If they agreed to be part of the study, they were asked to fill a brief sociolinguistic questionnaire (Li et al., 2020).They were then randomly assigned to one of four experimental conditions: low conflict description in NL-54 participants; low conflict description in FL-53 participants; high conflict description in NL-46 participants; and high conflict description in FL-64 participants..

All participants were then showed the introductory text, followed by the description corresponding to their experimental group. They were then asked to report how likely it was that the individual (who they had just read about) belonged to the group of people who had an injury in the last year, i.e., to give the estimation of probability: "*Please indicate your probability that the person described belongs to the injured group, on a scale from 0 to 100."* After indicating their choice, participants were asked to estimate the level of confidence in their answer using the confidence and bet scales.

To conclude the experimental task, participants were required to complete an additional task that measures the level of proficiency in English (i.e., LexTALE).

Measures and analyses

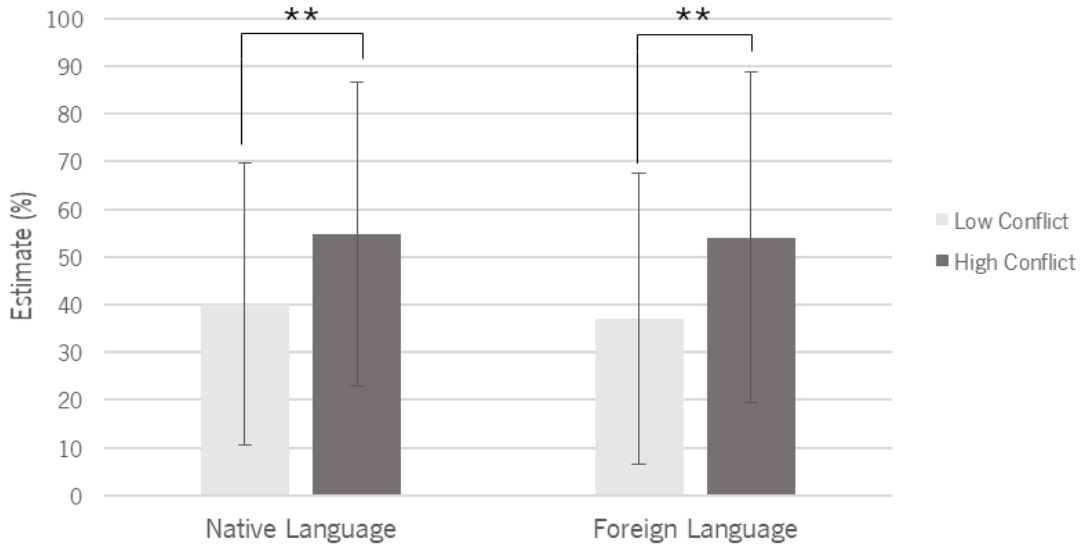
Three variables were initially measured, average *estimated probability*, *confidence rating* and *bet size*. To maintain a comparable scale with the *confidence rating*, the *bet size* was transformed using a logarithmic base 2 function. A set of multiple linear regression models were constructed to investigate the effect of language (foreign and native) and conflict (low and high), as well as their interactions, on the prediction of various dependent variables, such as *estimated probability*, *confidence rating* and *bet size*. Several variables were included as covariates in each model: age of the participant, age of acquisition of the foreign language (AoA), mean self-reported proficiency in the foreign language, and LexTALE's score. Additionally, we also decided to construct a couple of *post-hoc* multiple logistic regression models on the triple interaction of language, conflict, and accuracy (1 = response following the base-rate, 0 = response against base rate), and on the triple interaction of language, conflict, and extreme scores. We considered an answer to be extreme when it was binary. I.e., the participant estimated probability was lower than 5% or higher than 95%).

All variables were introduced in the models; no method was used for variable selection. Continuous variables were centered and transformed into Z-scores to standardize them. Following the guidelines of Schad et al. (2020), dichotomous variables were coded using sum contrast coding, with a value of -0.5 for the first level and +0.5 for the second level of each factor. The multicollinearity of the variables was also examined and found to be non-existent with all VIF values less than 3 (Zuur et al., 2010).

Results

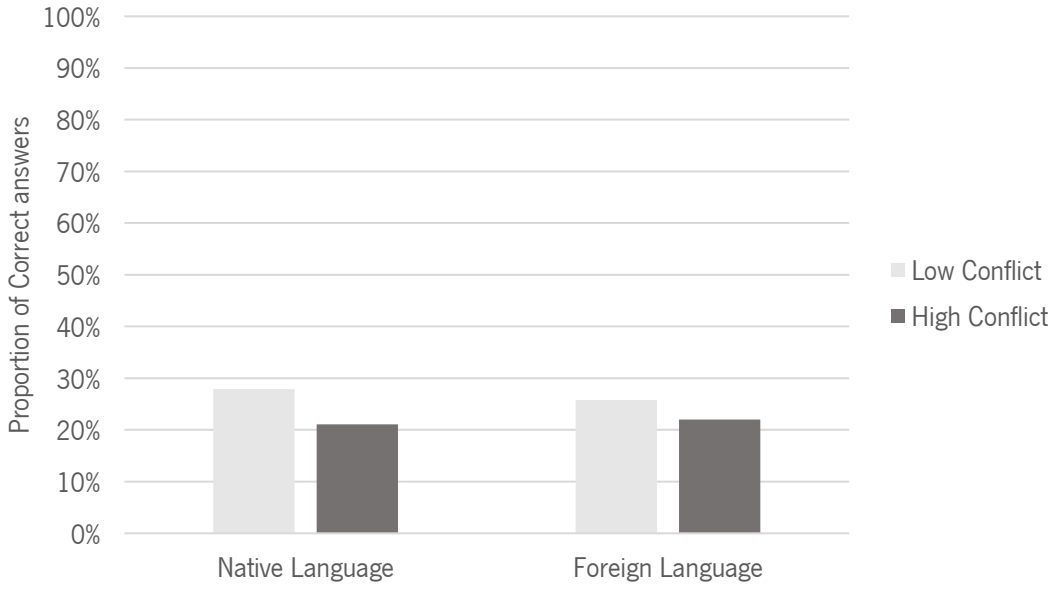
For native language speakers, the mean probability estimate for low conflict was 40.12 (SD = 29.7) and for high conflict was 54.87 (SD = 31.8). For foreign language speakers, the mean probability estimate for low conflict was 37.16 (SD = 30.5) and for high conflict was 54.02 (SD = 34.7), see figure 1.

Figure 1. Average estimated probability for the different languages and conflict levels. ** $p < .01$



The proportion of accurate answers when using the NL for low conflict was 28%, while the mean accuracy for high conflict was 21%. In the FL the proportion of accurate answers for low conflict was 26%, while the mean accuracy for high conflict was 22%. However, accuracy was generally low for both types of scenarios, with a wider range of scores for high conflict scenarios, see figure 2.

Figure 2. Proportion of correct answers for the different languages and conflict levels.



Regarding extreme answers, the proportion in the NL condition with low conflict was 23%, while the proportion of extreme answers for high conflict was 24%. In the FL, the proportion of extreme answers for low conflict was 29%, while the mean accuracy for high conflict was 38%, see figure 3.

Figure 3. Proportion of extreme answers for the different languages and conflict levels.

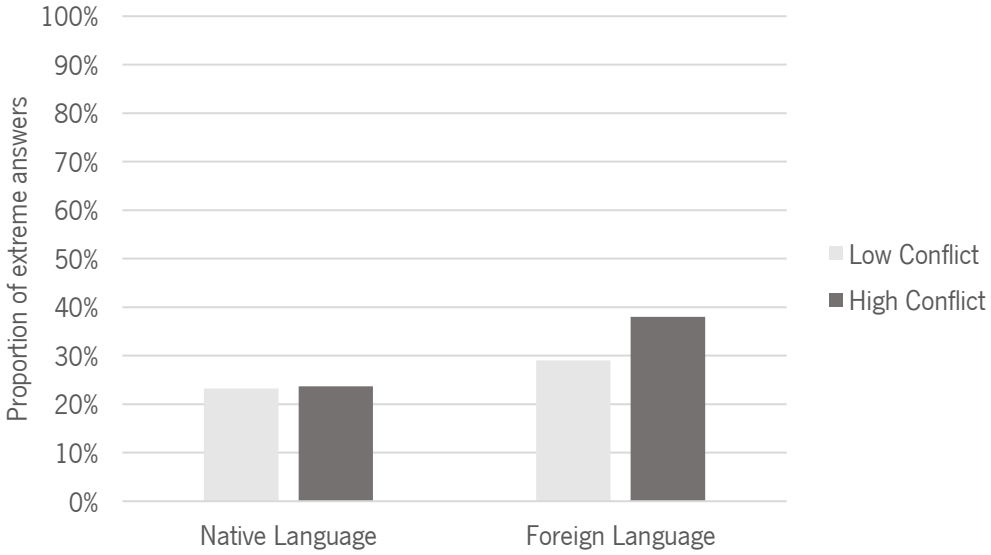


Table 1 also presents the summarized data on probability estimates, accuracy and extremes.

Table 1. The data on probability estimates are presented as means (standard deviations). ** $p < .01$

Language	Conflict level	*probability estimates*		Δ	Accuracy	Extremes
Native	**Low**	40.12	(29.7)	14.75**	28%	23%
	High	54.87	(31.8)		21%	24%
Foreign	**Low**	37.16	(30.5)	16.86**	26%	29%
	High	54.02	(34.7)		22%	38%

Regarding the probabilities estimated by the participants, we found a significant emotion effect, i.e., the probability estimate of the "high conflict" condition was higher than that of the "low conflict" condition (Emotion: $\beta = -15.62$, SE = 5.22, t = -3.00, $p = 0.003$). But no significant language main effect was found (Language: $\beta = 1.07$, SE = 5.22, t = 0.20, $p = 0.838$) nor interaction effect between language and emotion (Language × Emotion interaction: $\beta = -3.03$, SE = 10.47, t = -0.29, $p = 0.773$). The model explains 8.5% of the variance of the dependent variable, as indicated by the coefficient of determination (R^2), adjusting to 4.3% after considering the number of predictors (adjusted R^2).

Additionally, no differences between NL and FL speakers in accuracy were found (Language : $\beta = 0.00$, SE = 0.07, t = 0.00, $p = 0.999$), nor any interaction between language and emotion (Language

× Emotion interaction: β = 0.04, SE = 0.14, t = 0.29, p = 0.772). Regarding extreme answers we also did not find significant effect of language, conflict or interaction between them (Conflict: β = -0.04, SE = 0.07, t = -0.52, p = 0.607); (Language: β = -0.11, SE = 0.07, t = -1.52, p = 0.130); (Language × conflict interaction: β = 0.10, SE = 0.15, t = 0.69, p = 0.491). We did find, however, a significant AoA effect on the likelihood of an extreme answer, indicating that participants with lower AoA in the second language gave more frequently extreme probability values than those with higher AoA (AoA: β = -0.08, SE = 0.04, t = -2.06, p = 0.041). The model explains 7.4% of the variance of the dependent variable, as indicated by the coefficient of determination (R^2), adjusting to 3.2% after considering the number of predictors (adjusted R^2).

Table 2 presents data on *confidence rating* and *bet size*. The data are presented as means with standard deviations in parentheses.

Table 2. Average confidence ratings and bet sizes.

Language	Conflict level	*confidence rating*	*bet size*
Native	**Low**	3.70 (0.86)	2.72 (1.92)
	High	3.63 (1.08)	2.29 (2.03)
Foreign	**Low**	3.77 (0.99)	3.06 (1.93)
	High	3.68 (1.02)	3.00 (2.04)

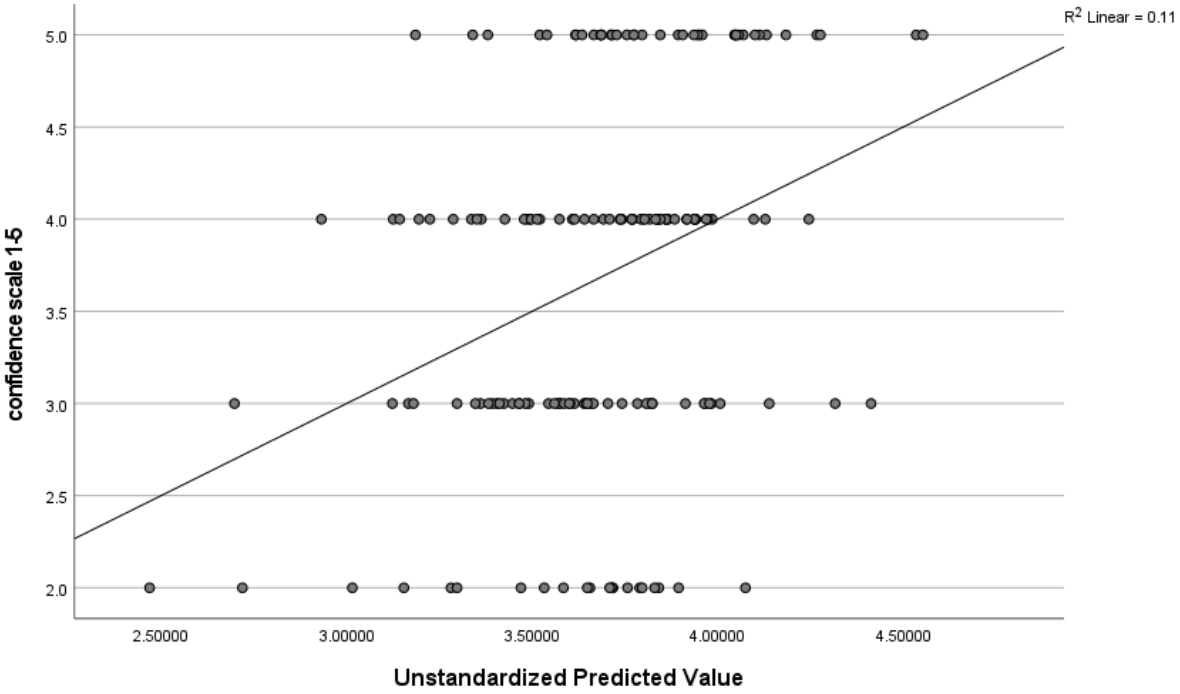
For native language speakers, the mean confidence rating for low conflict was 3.70 (SD = 0.86) and for high conflict was 3.63 (SD = 1.08). The mean bet size for low conflict was 2.72 (SD = 1.92), and for high conflict was 2.29 (SD = 2.03). For foreign language speakers, the mean confidence rating for low conflict was 3.77 (SD = 0.99) and for high conflict was 3.68 (SD = 1.02). The mean bet size for low conflict was 3.06 (SD = 1.93), and for high conflict was 3.00 (SD = 2.04).

Overall, the results suggest that there were no significant differences in confidence ratings between native and foreign language speakers for either low or high conflict scenarios. However, foreign language speakers tended to bet more than native language speakers in both low and high conflict scenarios. It is important to note that the bet sizes had a marginally wider range of values for high conflict scenarios, indicating that participants may have had greater uncertainty or risk-taking behaviour in these situations.

We found significant effects of age, AoA, and foreign proficiency on confidence, though. The model explains 11.2% of the variance of the dependent variable, as indicated by the coefficient of

determination (R^2), adjusting to 7.2% after considering the number of predictors (adjusted R^2), see figure 4.

Figure 4. Multiple linear regression model for confidence



Specifically, confidence increased linearly with age (Age: β = 0.21, SE = 0.08, t = 2.65, p = 0.009), confidence decreased linearly with AoA of the second language (AoA: β = -0.17, SE = 0.08, t = -2.03, p = 0.044), and confidence increased linearly with proficiency in the second language (Foreign Proficiency: β = 0.22, SE = 0.08, t = 2.59, p = 0.010).

Additionally, there was a significant AoA effect on bet, indicating that bet sizes decreased linearly with AoA of the second language (AoA: β = -0.32, SE = 0.16, t = -1.98, p = 0.050), while a significant age effect on bet showed that bet increased linearly with age (Age: β = 0.57, SE = 0.16, t = 3.64, p < 0.001). In this case the model explains 13.9% of the variance of the dependent variable, as indicated by the coefficient of determination (R^2), adjusting to 10% after considering the number of predictors (adjusted R^2).

Discussion

The reduced emotionality account of the FLe was tested through a modified base-rate task. In the task, participants were asked to estimate the probability of a person belonging to a particular group while ignoring the stereotypical information they have about the person. We found no significative

differences in accuracy, nor in the average size of overestimations between NL and FL speakers, thus not supporting either model proposed by (Białek et al., 2020).

Our lack of significative results might stem from missing a critical assumption made by the CTM: "This Conflict Threshold Model would predict that when experiencing equivalent conflict between competing intuitions, NL reasoners will detect a conflict and FL reasoners will not." (Białek et al., 2020, p.12). This poses the problem of how to balance the subjective weight of some intuitions, that we are still trying to measure. It has been consistently shown that not all intuitions will have the same subjective weight. In particular, intuitions based on congruency between inputs that induce a sense of cognitive ease, like an irrelevant description being evocative of a given stereotype, seem to outweigh logical or probabilistic intuitions (i.e., a base-rate), even when explicitly stated (Kahneman, 2003, 2012; Kahneman & Tversky, 1973; Vives et al., 2018). So how should we go about making them equivalent? Vives et al. (2018) had three different conditions that ranged from congruent to incongruent, from conflict to facilitation. But as their results have shown, not even when the conflict was highest were any of the participants, regardless of NL or FL, able to detect it. This means that even then the heuristic intuition was stronger than the probabilistic or statistical one.

The question/answer congruency vs the answer/answer congruency.

Because we intended to produce two equivalent intuitions, we focused on stimuli that always produced some level of conflict. However, as we see it, our design might potentially have one big confounding variable. There is more conflict between the question being asked with one of the intuitions than the other. I.e., the question being asked "*Please indicate your probability that the person described belongs to the injured group, on a scale from 0 to 100*" is congruent with the description whose gist is "he has been injured in the past" but not with the base-rate, with a lower probability of belonging to the injured group. Vives et al. (2018) avoided this problem by asking a dichotomous question that did not point towards any of the possible answers, but given that we were interested in the size of the overestimations, that was not possible in our experiment. In hindsight we might have better controlled for this by also asking the question in the negative form, i.e., how likely is he <u>not</u> to belong to the injured group and reversing the base-rate. This would make it so that the inputs description and base-rate would be incongruent with the category in the question, i.e., the base-rate would point towards injured: thirty not injured vs. seventy injured; and the description would also point towards the injured group. Even though these inputs would now be congruent with one another, any overestimation can only be attributable to the representativeness bias elicited by the description. Which in turn would allow us to better gauge the strength of the two competing intuitions (heuristic vs. statistical). The major setback of

that approach being, that the more complex the design is the bigger the sample required, and we were limited by the availability of subjects and time.

Aside from said limitations, the results of our study are in line with previous work that studied the possibility of a FLe when performing base-rate tasks (Vives et al., 2018). However, our results should be taken carefully since the FLe is a fickle effect (Leivada et al., 2021). It does not only depend on the task, the stimuli, and the proficiency level of the participants, but also on the languages chosen. Just as no two languages are completely identical, no two pairings of languages are identical as well. What this implies is that assuming everything else is identical, there is still a possibility that no FLe will be found due to the characteristics of the two languages – social, cultural, formal, etc. (Dylman & Champoux-Larsson, 2020; Miozzo et al., 2020).

Regarding our measures of confidence, we couldn't perform all the analysis we intended due to the low number of participants who were accurate in their estimations. However, our analysis showed that the earlier participants acquired their second language (AoA), the more likely they were to report higher levels of confidence, place higher bets and the more likely they were to give an extreme answer. The confidence rating was also affected by the age of the subjects. Confidence increased linearly with age, and proficiency level with the FL. There is some evidence to suggest that people may become more confident with age, although the relationship between age and confidence is complex and may be influenced by a variety of factors. One possible reason why people may become more confident with age is that they accumulate more life experience over time, which can help them to develop a greater sense of self-efficacy and resilience. It's also worth noting that confidence can be influenced by a variety of individual factors such as personality traits, cognitive abilities, and self-esteem, as well as by social and cultural factors such as gender roles and societal expectations.

It is also interesting to note that AoA seems to drive some of the variability that we see on our data regarding confidence and the probability of an extreme answer despite the task being performed in either the native or the foreign language. This might be just a reflection of some subjective characteristics of our participants, maybe those who are generally more confident started to learn a second language sooner. It is possible though, that it is also the other way round, people who start learning a second language sooner become gradually more confident over time. This hypothesis might be connected to the ideas presented in several language models that postulate different language nodes that actively compete for cognitive resources. The BIA+ model, for example, assumes that bilinguals have a shared lexicon for their two languages and that the activation of a word in on1e language can spread to its translation equivalent in the other language. This activation is influenced by the level of proficiency of the bilingual in

each language, as well as the degree of overlap between the phonological, orthographic, and semantic representations of the words in each language (Dijkstra & van Heuven, 2002). Therefore, the age at which the second language is acquired can also be a factor. If the second language is learned earlier in life, it may have a more profound effect on the individual's cognitive processing and language abilities. That is besides the scope of the present work, however, and our study does not allow us to make that distinction.

Despite the limitations of Vives et al. (2018)'s study and our own, the results taken together seem to point towards the non-existence of a FLe in base-rate tasks. However, if a base-rate task was to be constructed in such a way that the strength of the competing intuitions was amenable to be measured and balanced, then an even more solid piece of evidence could be available.

References

Białek, M., Muda, R., Stewart, K., Niszczota, P., & Pieńkosz, D. (2020). Thinking in a foreign language distorts allocation of cognitive effort: Evidence from reasoning. *Cognition*, *205*(October), 104420. https://doi.org/10.1016/j.cognition.2020.104420

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral Foreign-Language Effect. Philosophical Psychology, 29(1), 23–40. https://doi.org/10.1080/09515089.2014.993063

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). "Piensa" twice: On the foreign language effect in decision making. *Cognition*, *130*(2), 236–254. https://doi.org/10.1016/j.cognition.2013.11.010

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0094842

Dijkstra, T., & van Heuven, W. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197. https://doi.org/10.1017/S1366728902003012

Dylman, A. S., & Champoux-Larsson, M. F. (2020). It's (not) all Greek to me: Boundaries of the foreign language effect. *Cognition*, *196*(November 2019). https://doi.org/10.1016/j.cognition.2019.104148

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Harris, C. L., Ayçiçeği, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics*, *24*(4), 561–579. https://doi.org/10.1017/S0142716403000286

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2012). *Pensar Depressa e Devagar* (4ª). Círculo de Leitores.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/https://doi.org/10.1037/h0034747

Kahneman, D., & Tversky, A. (1979). Prospect Theory : An Analysis of Decision under Risk Linked references are available on JSTOR for this article : All use subject to http://about.jstor.org/terms E C O N OMETRICA I C I. *Source: Econometrica*, *47*(2), 263–292. http://www.jstor.org/stable/1914185%5Cnhttp://www.jstor.org/%5Cnhttp://www.jstor.org/action/showPublisher?publisherCode=econosoc.%5Cnhttp://www.jstor.org

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases. *Psychological Science*, *23*(6), 661–668. https://doi.org/10.1177/0956797611432178

Leivada, E., Westergaard, M., Duñabeitia, J. A., & Rothman, J. (2021). On the phantom-like appearance of bilingualism effects on neurocognition: (How) should we proceed? *Bilingualism: Language and Cognition*, *24*(1), 197–210. https://doi.org/DOI: 10.1017/S1366728920000358

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*, 938–944. https://doi.org/10.1017/S1366728918001153

Lowe, C. J., Cho, I., Goldsmith, S. F., & Morton, J. B. (2021). The Bilingual Advantage in Children's Executive Functioning Is Not Related to Language Status: A Meta-Analytic Review. *Psychological Science*, *32*(7), 1115–1146. https://doi.org/10.1177/0956797621993108

Miozzo, M., Navarrete, E., Ongis, M., Mello, E., Girotto, V., & Peressotti, F. (2020). Foreign language effect in decision-making: How foreign is it? *Cognition*, *199*(February). https://doi.org/10.1016/j.cognition.2020.104245

Rosselli, M., Velez uribe, I., & Ardila, A. (2017). *Emotional Associations of Words in L1 and L2 in Bilinguals* (pp. 39–72). Springer International Publishing. https://doi.org/10.1007/978-3-319-64099-0_3

Vives, M. L., Aparici, M., & Costa, A. (2018). The limits of the foreign language effect on decision-making: The case of the outcome bias and the representativeness heuristic. *PLoS ONE*, *13*(9), 1–14. https://doi.org/10.1371/journal.pone.0203528

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common

statistical problems. Methods in Ecology and Evolution, 1(1), 3–14. doi: 10.1111/j.2041-210X.2009.00001.x