THE FOREIGN LANGUAGE EFFECT ON THE ILLUSION OF CAUSALITY:

A replication attempt and an explorative analysis of the mechanisms

ABSTRACT

Díaz-Lago and Matute (2019b) supported in a series of two experiments that a cognitive bias, the illusion of causality, can be attenuated when the task eliciting this illusion is presented in a foreign language. This initial result in coherent with a certain set of literature (e.g., Circi et al., 2021; Del Maschio et al., 2022; Purpuri et al., 2024), which over the years has shown that the usage of a foreign language can influence decision-making in tasks commonly used to assess flawed reasoning processes (i.e., the foreign language effect; FLE).

The present study aims to achieve two primary objectives. First, we seek to replicate the original findings of Díaz-Lago and Matute (2019a) with a more nuanced approach, utilizing suitable statistical inference tools to assess whether we can directly support the presence or absence of the effect, while also estimating the ES for the observed reduction in the illusion of causality. The second goal of our study is to provide an exploratory framework that may help identify which explanation(s), or combination thereof, could account for the FLE observed in the context of the illusion of causality.

1 A BIAS REDUCED VIA A FOREIGN LANGUAGE

1.1 Illusion of Causality

The illusion of causality is a bias where individuals mistakenly perceive the presence of a causal connection between unrelated events in associative learning contexts, where the cause and effect presence or absence can appear with different frequencies (Matute et al., 2022). People tend to overestimate a causal link between a potential cause and an outcome after reviewing a series of trials, each characterized by the presence or absence of the potential cause and the outcome. The illusion of causality can be understood as biased evaluation of raw data, where individuals tend to prioritize true positives – scenarios where both the supposed cause and effect are observed – over true negatives, false positives, and false negatives (Matute et al., 2015). For that reason this bias can be included in the broader category of inductive type of reasoning, as people have to draw general conclusions based on the evaluation of a series of trials, and its belonging to the statistical bias can be justified considering the underlying processes for which our associative mechanism tends to overweight the cumulative evidence proving the existence of an effect rather than checking for a disconfirmation.

One of the most commonly used paradigms for studying causal learning and the illusion of causality is the Contingency Learning Task (CLT). In this paradigm, participants are presented with a series of trials, each involving the presence or absence of two events: A (the potential cause) and B (the supposed effect). The combination of these events' presence or absence leads to four possible trial types: (a) both event A and event B occur (i.e., the cause and effect co-occur), (b) only event A occurs (i.e., the cause is present without the effect), (c) only event B occurs (i.e., the effect occurs without the cause), and (d) neither event A nor event B occurs (i.e., neither the cause nor the effect is present). These scenarios may appear with varying frequencies, ranging from zero on. Thus, each trial presents evidence for the relationship between potential cause A and effect B, with event A preceding event B in all cases. This design ensures that only event A can signal event B, thereby suggesting a potential causal link. Typically, the events A and B are chosen to be plausibly causally related. After completing a number of trials (the exact number is determined by the researcher), participants are asked to rate the perceived degree of causal connection between the events. A numerical scale from 0 to 100 is commonly used, where 0 indicates no perceived causal connection and 100 represents the strongest possible causal link.

Researchers manipulate the presence or absence of events A and B by adjusting their proportions, thus controlling the statistical relationship between the events. The frequencies of the four possible scenarios (a, b, c, d) are systematically varied to create different levels of contingency between the potential cause and the supposed effect. For instance, in a true contingency condition, the cue and outcome are frequently observed together (i.e., a high frequency of scenario a) and rarely occur independently (i.e., low frequencies for scenarios b and c). In contrast, in a null contingency condition, the cue and the outcome occur independently, leading to more balanced or low frequencies across all four scenarios.

In certain experimental conditions where no real causal link exists, people tend to overestimate the strength of the causal connection between the events. This overestimation is considered the operational definition of the illusion of causality. In such cases, participants are exposed to a series of trials where normative indices (e.g., the ΔP) indicate no causal relationship (null contingency condition); however, by increasing the frequencies of the cause, effect, or both, while maintaining independence between events, researchers can create the illusion that a causal link is present (i.e., effectively, inducing the illusion of causality – null contingency illusory condition).

1.2 Foreign Language Effect

Recent interest in this field has explored whether the illusion of causality can be reduced. Examining the contexts in which this illusion can be diminished not only offers a meaningful goal for educating the public to overcome the illusion but also helps identify the variables influencing the underlying processes, thereby enhancing our understanding of the phenomenon. In this perspective, Díaz-Lago and Matute (2019b) found that using a foreign language could reduce the manifestation of the illusion of causality.

The Foreign Language Effect (FLE) was first described by Keysar et al. (2012), who found that participants exposed to a decision-making task in a foreign language (FL) exhibited less biased responses compared to those performing the task in their native language (NL). In a series of six experiments, the authors concluded that the use of a FL can lead to a reduction in decision-making biases; for example, when the Asian Disease Problem was presented in a relatively low-proficiency FL, the effect of framing options in terms of gains or losses was diminished compared to the NL condition, and participants tended to choose the risk-averse option to a similar extent in both the gain and losses conditions. Essentially, Keysar et al. (2012) argued that, in a FL, people are less risk-averse and tend to weigh negative outcomes less heavily than positive ones, thereby making more rational decisions.

Over the years, the FLE has been found to be consistent across various tasks, including loss-aversion paradigms, decision-making, and moral dilemmas (Circi et al., 2021). For instance, Costa, Foucart, Arnon, et al. (2014) replicated the findings of Keysar et al. (2012), extending the evidence of the phenomenon to other heuristics and supporting, once again, that decision-making when problems are presented in a FL, is less susceptible to biases. In the area of moral dilemmas, which have been extensively explored with respect to the FLE, the use of a FL appears to promote more utilitarian responses (e.g., Costa, Foucart, Hayakawa, et al., 2014). These responses are considered more rational as they prioritize maximizing outcomes.

1.3 FLE in the illusion of causality

The reduction of the illusion of causality through the use of a FL (Díaz-Lago & Matute, 2019b) is an intriguing preliminary finding, as it could provide valuable insights into this cognitive bias. In a set of two consecutive experiments, Díaz-Lago and Matute (2019b) did find that a FL could reduce the illusion of causality. Their main results were explained by the idea that cognitive disfluency (i.e., the difficulty on the task), provoked by a FL, promotes greater use of analytical thinking, leading to more normative responses. However the results of a recent study (Dalla Bona & Vicovaro, 2024) challenge the idea that cognitive disfluency reduces the illusion of causality. In contrast to Díaz-Lago and Matute (2019a) conclusions, which suggest that cognitive control is enhanced through process (dis)fluency – where the manipulated difficulty on the task is believed to act as a metacognitive cue that fosters more systematic and analytical thinking – Dalla Bona and Vicovaro (2024) found no evidence that the increased difficulty, manipulated via the perceptual features of stimuli in the associative learning task, influences the strength of the illusion of causality. This highlights the need for further research into the role of FL in causal reasoning, to determine whether and how FL manipulations can influence the illusion of causality. Therefore, a replication of the original study is needed, with an explorative structure able to indicate what explanation could underlie the reduction of the illusion of causality.

2 REPLICATION ATTEMPT

2.1 Original study

Díaz-Lago and Matute (2019b) conducted two experiments (N = 36 in the first experiment, N = 80 in the second) in which they observed a reduction in the illusion of causality when a FL was used. We reanalyzed their publicly available data set on Open Science Framework (OSF) to derive general insights for conducting a replication study (the reanalysis file is available at: https://osf.io/hvgkx/?view_only=7098178875224cf3b0b6890b209432ea). In our reanalysis of their original experiments, we recomputed the effect size between the two groups (i.e., NL group and FL group) in both experiments. The first experiment yielded a Cohen's d of 1.4 (a very large effect), while the second experiment yielded a Cohen's d of 0.8 (a large effect) for the difference between the two groups. When aggregating the results from both experiments, the overall effect size for the difference between the NL and FL groups was Cohen's d = 1. We decided to focus on the second experiment because its sample characteristics are more similar to those of the final sample we plan to collect. For instance, the first experiment involved English-speaking Erasmus students who began learning Spanish as a second language later in life (Age of Acquisition of a FL; AoA: M = 12.61, SD = 4.08) compared to participants in the second experiment which involved Spanish speakers who learned English as a second language (AoA: M = 6.57, SD = 4.32). Our experiment aims to collect data from native Italian speakers who learned English as a second language in school, making the second experiment of Díaz-Lago and Matute (2019b) a more relevant comparison.

It is worth highlighting that the effect sizes found in Díaz-Lago and Matute (2019b) experiments are very large, compared with the small effect size of Cohens d = 0.2 that emerged in recent meta-analyses on the FLE (e.g., Del Maschio et al., 2022; Circi et al., 2021). However, although it cannot be excluded that the effect size observed by Díaz-Lago and Matute (2019b) might be an overestimate due to the relatively small sample size, caution must be exercised before drawing direct conclusions based solely on meta-analytic data. Indeed, these meta-analyses primarily analyzed the FLE in moral dilemma tasks, which involve different decision-making processes compared to associative learning tasks. Furthermore, the core characteristics of the response variable on which effect sizes are computed clearly differ across the two tasks, as in moral dilemmas, participants are asked to choose between two or three options, whereas in the associative learning task participants are asked to rate the degree of causal connection between two events on a Likert scale from 0 to 100.

2.2 Rationale for a replication

In our experiment, we aim to conduct a two-group comparison between participants (native Italians with moderate/low English fluency) who complete the experiment under a null contingency illusory condition in either Italian (i.e., NL) or English (i.e., FL). The main objective is to investigate potential differences between these two groups in terms of the degree of the illusion of causality.

In order to establish the sample size required for testing the hypothesis of a difference between the two groups (one in Italian and one in English) with respect to the illusion of causality, we employed a simulative approach. Specifically, we aimed to simulate data to model a difference between the two groups on a Likert scale ranging from 0 to 100 (the simulation file is available at: https://osf.io/hvgkx/?view_only=7098178875224cf3b0b6890b209432ea).

We based our simulation on a distribution analysis of data from the study by Dalla Bona and Vicovaro (2024), ultimately concluding that the NL groups causality evaluation variable follows a truncated normal distribution (given the bounded nature of the scale). The mean for this group was hypothesized to range between 55 and 65, with a standard deviation of 20. For the FL group, we hypothesized a slightly higher standard deviation, following the findings of Díaz-Lago and Matute (2019b) (i.e., between 20 and 25).

Regarding the effect size between the two groups, we considered a range of possible mean differences, using a heuristic criterion based on the observed tendency of participants to anchor their responses on numbers ending in 5 or 10. We hypothesized that the minimal reduction in the illusion would be of 5 points, which would suggest that participants tend to anchor their responses to the previous "anchor" point. This minimum shift aligns with findings from meta-analyses showing that the FLE, on average, yields a Cohens *d* of 0.2. The maximum plausible shift in causality evaluation, according to Díaz-Lago and Matute (2019b), was hypothesized to be four anchors, or 20 points.

Given that the means are theoretical values and that the true effect size remains uncertain, but is expected to lie somewhere between these two extremes, we simulated mean differences across a range from 5 to 20 points, accounting for the anticipated increase in variability for the FL group and the uncertainty regarding the true mean for the NL group. Additionally, we considered the bounded and discrete nature of the Likert scale, where responses can only be integers.

From these simulations, we generated a prior distribution for Cohen's *d*, which was then incorporated as the design and analysis prior into a Bayesian Factor Design Analysis.

Design analysis using a Bayes Factor approach for the comparison between the two groups showed that, to achieve the required power of 0.8 (i.e., a minimum 80% of simulated studies yielding results in the correct direction for 5000 simulations), we would need to collect 110 participants per group (the Bayes Factor Design Analysis file is available at: https://osf.io/hvgkx/?view_only=7098178875224cf3b0b6890b209432ea).

3 EXPLORATORY ANALYSIS

3.1 Rationale

In addition to attempting to replicate the original findings by Díaz-Lago and Matute (2019b), we aim to incorporate appropriate measures to help identify plausible explanations for any observed reduction in the illusion of causality due to the use of a FL. Three main hypotheses for the FLE have been proposed over the years (Del Maschio et al., 2022).

3.2 Enhanced cognitive control

A possible explanation for the observed reduction in the illusion of causality – an effect that should not be influenced by emotional factors – when using a FL, is that the FL serves as a metacognitive cue, signaling to participants that the task might be challenging (Alter et al., 2007). This perception of difficulty could lead to a more focused state of mind, shifting cognitive processing from System 1 (intuitive and automatic thinking) to System 2 (deliberate and analytical thinking), as described in dual-process models of cognition. The use of the FL may thus create an expectation of difficulty, inducing participants to pay closer attention in order to "solve" the task.

To provide evidence for this hypothesis, we plan to utilize a scale measuring perceived task difficulty in order to examine whether the reduction in the illusion of causality is mediated by the construct of disfluency. Specifically, disfluency can be conceptualized as a unipolar dimension ranging from fluent to disfluent (or, equivalently, from effortless to effortful). Given this, a single item is considered sufficient to capture participants' perceived difficulty (as shown by Graf et al., 2018):

- "How difficult did you find the reading and comprehension activities during the task?" (7-point Likert scale)

In addition to this subjective evaluation of task difficulty, we measure an objective index of task difficulty: the total time taken to complete the experiment. This index will be used to assess, through a comparison between the two groups, whether there is a difference in time spent on the task.

If a reduction in the illusion of causality is observed, and the enhanced cognitive control theory provides a valid theoretical explanation for this reduction, we would expect to see that the item on subjective task difficulty mediates the relationship between language and the observed reduction in the illusion.

3.3 Emotional distance

Another possible explanation, which has been a leading theory for the FLE for several years, suggests that a FL could create greater psychological distance from the affective system, or that a NL might elicit a stronger emotional response (Del Maschio et al., 2022). This would imply that a reduction in biases may only occur in emotionally charged tasks (such as moral dilemmas), as has already been supported by some research. This led us to question whether there is an emotional component underlying the illusion of causality paradigm.

Typically, the illusion of causality has been studied using a paradigm in which participants evaluate the causal connection between a potential cause and its effect within a cover story. For example, in the Allergy task (Matute et al., 2015), which is commonly used to assess the illusion of causality and corresponds to the task used in our present study, participants are asked to imagine themselves as doctors and assess whether a fictitious medicine can cure a fictitious disease by observing a series of patients suffering from this disease. While at first glance this task might not seem particularly emotionally engaging, there are several aspects of the cover story that warrant further consideration. It has already been supported that motivational variables, manipulated through the cover story with which the illusion is elicited, can enhance the illusion of causality (Matute et al., 2022). The allergy task may have an emotional component that enhances the illusion: participants might perceive the disease task as emotionally engaging, which could increase their motivation to see the medicine as effective, thereby amplifying their susceptibility to the causality bias. The use of a FL, in this case, may reduce emotional engagement with the task, leading to a decrease in the illusion of causality for participants who complete the task in a FL compared to those who do so in a NL.

To explore the hypothesis of a reduction in the emotional component, we plan to introduce a scale that measures participants' emotional engagement with the task. Specifically, we will use the Affective Slider (Betella & Verschure, 2016), the digital version of the more well-known Self-assessment Manikin (SAM; Bynion and Feldner, 2020), a non-verbal pictorial scale typically employed to quickly assess participants' subjective emotional responses to a certain event (Bradley & Lang, 1994).

The original SAM scale (Bynion & Feldner, 2020) included a measure of emotional valence (ranging from a smiling figure to a frowning figure; 5-point Likert scale), a measure of emotional intensity (ranging from an alert figure to a sleeping figure; 5-point Likert scale), and a measure of dominance (ranging from a small figure to a big figure; 5-point Likert scale). In contrast, the Affective Slider only includes standardized measures of valence and arousal on an analog scale from 0 to 1 (with 0.5 assumed to be the mid-point). According to the hypothesis that the FLE is mediated by increased emotional distance from the task at hand, a difference between the two groups in emotional intensity ratings is expected, with the FL group reporting lower levels of intensity, as we can interpret the intensity as a measure of emotional activation provoked by the task.

Furthermore, we hypothesize that this difference in emotional intensity will mediate the causal judgment. The valence scale will be used to control for potential differences in the types of emotions elicited by the task, specifically to examine whether the fictitious story elicits more negative emotions in the NL group compared to the FL group. The two analog scales will be presented after the instructions of the Allergy task, in which participants have to identify as a doctor and are instructed to seek a potential relationship between the medicine and the patients' healing. The rationale for including these scales right after the instructions is that we posit only the instructions to be emotionally charged, whereas the main associative task could perhaps induce boredom in participants, as it is a very repetitive task. The two questions will be presented as follows:

- "How did you feel while reading this story that was presented to you? The first slider asks which emotion you are feeling (from sad to happy)."
- "How did you feel while reading this story that was presented to you? The second slider asks how activated you feel (from calm to activated)."

3.4 Context of acquisition and acquired norms

Geipel et al. (2015) suggested that the FLE in the context of moral dilemmas could be due to the context in which a NL and FL are acquired. As memory systems have been shown to include a trace of the language of encoding, a moral dilemma presented in NL may trigger stronger language-dependent access to sociocultural and moral norms than in FL. While the connection between the illusion of causality and this explanation is not immediately given, we can attempt to frame a theoretical rationale.

We foresee two ways in which language encoding could influence associative learning in the context of the illusion of causality. First, prior to the experiment, where participants are asked to evaluate the association between a fictitious medicine and disease healing – an association that is learned rather than innate, but assumed to be familiar – using the NL may facilitate access to prior memories where medicines have been seen as effective treatments. In this sense, there is existing knowledge, more easily accessed in the NL, that strengthens the connection between the medicine and healing before the associative trials begin. In contrast, when using a FL, these associations may be harder to retrieve, resulting in a weaker connection between the medicine and the healing process compared to the NL.

However, this mechanism seems to conflict with the results of the original experiment, where the average associative strength between the medicine and healing was consistent across control groups in the true contingency scenario (i.e., where a genuine causal relationship exists), regardless of the language used. Its also unlikely that this discrepancy is due to a ceiling effect, as participants' ratings of the causal connection were 66.1 (SD = 11.39) for the NL group and 68.75 (SD = 19.32) for the FL group, out of a maximum of 100.

To examine whether prior expectations are stronger in the NL group compared to the FL group, we plan to introduce a question that assesses participants' pre-existing beliefs. Specifically, we will ask them to rate the extent to which they believe there is a causal link between medicine and healing, based on their personal experiences. However, we must be mindful of potential bias, as responses to this expectation question could influence participants' evaluations in the main associative learning task, and vice versa. To mitigate this risk, we believe that if one measure has to be influenced by prior responses, it is preferable to have the expectation question be the one affected, in order to preserve the integrity of the main dependent variable.

Therefore, we propose asking the expectation question after the main evaluation of the causal connection. The question will be as follows:

- "Beyond the specific case of Batatrim, in your PERSONAL experience, how effective are medications in treating diseases IN GENERAL?" (Rated on a 101-point Likert scale from 0 = "Definitely not" to 100 = "Definitely yes").

If a difference is observed between the NL and FL groups, and this difference correlates linearly with the main evaluation of the causal connection, it would suggest that expectations play a role in diminishing the illusion of causality caused by the FL condition.

An alternative explanation is that the FLE may impact the associative learning process at a more mechanistic level, particularly if semantic associations are weaker in the FL condition (Kroll & Stewart, 1994). This could lead to slower associative learning and weaker associations after the learning trials. Weaker associative learning may degrade more quickly, meaning that participants in the FL group could experience greater difficulty estimating the number of instances they encountered during the associative learning trials.

However, testing this explanation within an exploratory framework is challenging, as it would require tracking how associations develop trial by trial, comparing how participants in the FL condition form weaker associations relative to those in the NL condition. Nonetherless to address the potential language encoding effect on learning and memory, we plan to introduce a follow-up question after the main associative task to explore this possible interaction more deeply. Participants will be asked to estimate the number of times they saw the following trial types:

- (a) The cause present with the effect
- (b) The cause present with the effect absent
- (c) The effect present with the cause absent
- (d) Neither the cause nor the effect present

We can hypothesize that participants in the FL group may underestimate the total number of trials, especially for trial type (a), where both the cause and effect are present. These questions will be asked after the inquiry regarding efficacy expectations.

This measure will also more generally help us assess whether there are differences in memory processes between the two groups due to language econding.

3.5 Measuring Language Features

An essential aspect of the experiment involves selecting an appropriate sample of Italian participants with intermediate proficiency in an unbalanced second language (i.e., English). Specifically, the recruitment form will target Italians with relatively intermediate proficiency in English. The form will clearly indicate that we are looking for participants who meet the following criteria:

- Are able to independently use the second language, as completing the task requires a certain level of proficiency in English;
- Do not hold a language certificate with a CEFR level higher than C1, which would indicate high levels of proficiency that could lead to invalid measurements;
- Began learning English in school and are not simultaneous bilinguals (i.e., individuals who learned both languages from birth);
- Do not use their second language on a daily basis (e.g., individuals living in an English-speaking country).

Participants who agree to the conditions outlined in the form will be randomly assigned to either the NL condition or the FL condition.

To assess language proficiency, we will rely on two criteria to ensure our measurements are meaningful for the experiment. First, some of the results should be comparable to those of the original study, particularly data regarding the Age of Acquisition (AoA) of the foreign language and proficiency levels across both languages. Additionally, we will introduce measures that are tailored to the specific characteristics of our sample (Italians with English as a second language). To this end, we have adapted several items from the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). In particular, we aim to measure the following in the Italian language:

- The possession of a language diploma with the question: "Do you possess a language diploma in English? If yes, please specify the type of diploma", with options ranging from A1 to C2. This scale aims to ensure that we do not have participants with high proficiency in English (i.e., C2 level).
- A pair of questions regarding whether Italian and English are the participants native languages: "Is English/Italian your native language?"
- The AoA of English and Italian with the question: "At what age did you begin learning English/Italian?" (adapted from the LEAP-Q). These questions will be presented only if the participants declared that the target language is NOT their native one.
- Self-assessment of proficiency in both NL and FL (if the participant has declared that this language is not their native): "Please rate your proficiency level in the following skills on a scale from 1 to 10" (separate ratings for NL and FL):
  - Speaking;
  - Reading;
  - Comprehension;
  - Writing.

  These questions will be presented only if the participants declared that the target language is NOT their native one.
- A question regarding whether the participant is currently studying (i.e., "Are you currently in a formal education program?) and a question regarding years of education: "How many years of formal education do you have?" These questions are included to check if there are any differences between the two NL and FL groups in terms of education.
- A question regarding the context of English usage: "Do you live in a country where English is the most widely spoken language?" (adapted from the LEAP-Q). This question aims to assess whether the participants live in a country where they speak English as a first language.

These questions will be presented at the end of the experiment in the NL for all participants. After this questionnaire, participants will take an English proficiency test (adapted from: https://www.cambridgeenglish.org/test-your-english/general-english) to objectively assess their English proficiency and provide a more detailed description of our sample.

REFERENCES

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. Journal of experimental psychology: General, 136(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

Betella, A., & Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. PLOS ONE, 11(2), 1–11. https://doi.org/10.1371/journal.pone.0148037

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bynion, T.-M., & Feldner, M. T. (2020). Self-assessment manikin. Encyclopedia of personality and individual differences, 4654–4656. https://doi.org/10.1007/978-3-319-24612-3_77

Circi, R., Gatti, D., Russo, V., & Vecchi, T. (2021). The foreign language effect on decision-making: A meta-analysis. Psychonomic Bulletin & Review, 28, 1131–1141. https://doi.org/10.3758/s13423-020-01871-z

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), e94842. https://doi.org/10.1371/journal.pone.0094842

Dalla Bona, S., & Vicovaro, M. (2024). Does perceptual disfluency affect the illusion of causality? Quarterly Journal of Experimental Psychology, 77(8), 1727–1744. https://doi.org/10.1177/17470218231220928

Del Maschio, N., Crespi, F., Peressotti, F., Abutalebi, J., & Sulpizio, S. (2022). Decision-making depends on language: A meta-analysis of the Foreign Language Effect. Bilingualism: Language and Cognition, 25(4), 617–630. https://doi.org/10.1017/s1366728921001012

Díaz-Lago, M., & Matute, H. (2019a). A hard to read font reduces the causality bias. Judgment and Decision Making, 14(5), 547–554. https://doi.org/10.1017/s1930297500004848

Díaz-Lago, M., & Matute, H. (2019b). Thinking in a foreign language reduces the causality bias. Quarterly Journal of Experimental Psychology, 72(1), 41–51. https://doi.org/10.1177/1747021818755326

Geipel, J., Hadjichristidis, C., & Surian, L. (2015). How foreign language shapes moral judgment. Journal of experimental social psychology, 59, 8–17. https://doi.org/10.1016/j.jesp.2015.02.001

Graf, L. K., Mayer, S., & Landwehr, J. R. (2018). Measuring processing fluency: One versus five items. Journal of Consumer Psychology, 28(3), 393–411. https://doi.org/10.1002/jcpy.1021

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

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. Journal of memory and language, 33(2), 149–174.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research, 50, 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Matute, H., Blanco, F., & Moreno-Fernández, M. M. (2022). Causality bias. In R. Pohl (Ed.), Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory (3rd ed., pp. 108–123). Routledge. https://doi.org/10.4324/9781003154730-9

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. Frontiers in psychology, 6, 888. https://doi.org/10.3389/fpsyg.2015.00888

Purpuri, S., Vasta, N., Filippi, R., Wei, L., & Mulatti, C. (2024). Does language shape the way we think? a review of the foreign language effect across domains. International Journal of Bilingualism, 13670069231225374. https://doi.org/10.1177/13670069231225374