

ARTICLE

Foreign to whom? Constraining the moral foreign language effect on bilinguals' language experience

Nicola Del Maschio¹, Gianpaolo Del Mauro¹, Camilla Bellini¹, Jubin Abutalebi^{1,2} and Simone Sulpizio^{3,4*}

¹Centre for Neurolinguistics and Psycholinguistics, Faculty of Psychology, Vita-Salute San Raffaele University; ²The Arctic University of Norway; ³Department of Psychology, University of Milano-Bicocca;

⁴Milan Center for Neuroscience (NeuroMI), University of Milano-Bicocca

*Corresponding author. E-mail: simone.sulpizio@unimib.it

(Received 12 April 2022; Revised 13 June 2022; Accepted 14 June 2022)

Abstract

The moral foreign language effect (MFLE) describes how people's decisions may change when a moral dilemma is presented in either their native (NL) or foreign language (FL). Growing attention is being directed to unpacking what aspects of bilingualism may influence the MFLE, though with mixed or inconclusive results. The current study aims to bridge this gap by adopting a conceptualization of bilingualism that frames this construct as a composite and continuous measure. In a between-group analysis, we asked 196 Italian–English bilinguals to perform a moral dilemmas task in either their NL (i.e., Italian) or FL (i.e., English). In a within-group analysis, we evaluated the effects of FL age of acquisition, FL proficiency, and language dominance – all measured as continuous variables – on moral decision-making. Overall, findings indicate that differences within bilinguals' language experience impact moral decisions in an FL. However, the effect of the linguistic factors considered was not ubiquitous across dilemmas, and not always emerged into a MFLE. In light of these results, our study addresses the importance of treating bilingualism as multidimensional, rather than a unitary variable. It also discusses the need to reconceptualize the FLE and its implications on moral decision-making.

Keywords: foreign language effect; bilingualism; moral decision-making; individual differences

1. Introduction

Imagine that you are driving through a busy city street when all of a sudden a young mother carrying a child trips and falls into the path of your vehicle. You are going too fast to brake in time; your only hope is to swerve out of the way. Unfortunately, the only place you can swerve is currently occupied by a little old lady. If you swerve to

avoid the young mother with her baby, you will seriously injure or kill the old lady. What will you do? Will you swerve and hit the old lady to avoid the mother with her child? (Adapted from Conway & Gawronski, 2013).

The *car accident* dilemma is an example of a class of moral conundrums utilized to study the psychological foundations of moral decision-making (Christensen & Gomila, 2012; Nichols & Mallon, 2006). Originally constructed as philosophical thought experiments, such dilemmas entail a difficult moral trade-off: deciding whether to cause harm to maximize overall outcomes. Rejecting a harmful action, independently of the circumstances and regardless of its outcomes, is supposed to reflect deontological inclinations that strictly align with moral norms (e.g., do not kill) (Kant, 1785/1959); endorsing that same action to promote a greater good is consistent with a utilitarian approach that privileges the maximization of aggregate welfare (Mill, 1861/1998). The individual and contextual factors that could influence a person's decision in dilemmas tailored to the deontological/utilitarian dichotomy have received substantial attention in the field of moral psychology (Bartels, 2008; Greene & Haidt, 2002; Körner et al., 2020). On the grounds of the literature on moral judgment, Christensen et al. (2014) have identified four main factors that seem to affect people's reactions to moral dilemmas and need to be controlled for in dilemma formulation: *Personal Force* (whether the agent is directly involved in the production of the harm, or not), *Benefit Recipient* (who gets the benefit), *Evitability* (whether the harm is avoidable, or not), and *Intentionality* (whether the harm is willed and used instrumentally, or a side-effect). It has been shown, for instance, that people are more likely to sacrifice an individual's life when the choice of action is impersonal rather than personal (e.g., Greene et al., 2009; Royzman & Baron, 2002), and that causing harm is more acceptable when it is produced as collateral damage than when it is the goal of an action (e.g., Borg et al., 2006; Cushman et al., 2006). In addition to conceptual differentiations between moral scenarios and individual difference variables, an initial investigation by Costa et al. (2014) and then several others suggested that decision outcomes may also be influenced by the language in which moral dilemmas are presented. In particular, systematically different choices have been reported when moral dilemmas are formulated in a native (NL) vs. a foreign language (FL), with a larger proportion of utilitarian (vs. deontological) decisions or judgments associated with FL contexts (e.g., Cipolletti et al., 2016; Costa et al., 2014; Geipel et al., 2015a). Crucially, while several studies have demonstrated a *Foreign Language Effect* on moral decision-making (MFLE), others have not been able to replicate this effect in contexts other than personal versions of the *Trolley* dilemma (e.g., Chan et al., 2016; Geipel et al., 2015a; Hayakawa et al., 2017), or have failed to report behavioral differences associated with FL contexts altogether (e.g., Białek et al., 2019; Brouwer, 2019; Čavar & Tytus, 2018).

The inconsistency of the above findings has inspired increasing interest in understanding the psychological and linguistic factors contributing to the FLE. A recent meta-analysis (Del Maschio et al., 2022) investigated the magnitude of the FLE on decision-making under conditions of risk (i.e., risk-aversion domain) and moral conflict (i.e., deontological/utilitarian dichotomy), and explored the potential effect of moderating variables related to second language experience (i.e., FL Age of Acquisition [AoA], FL proficiency) and experimental design (i.e., Problem type, Personal force, Task modality). Although the study reported a reliable FLE on decision outcomes, it was not able to detect a contribution of moderator variables on the observed effect. A second meta-analytic study by Stankovic et al. (2022)

restricted the literature search to the moral decision-making domain, and analyzed whether self-reported FL proficiency (reading, writing, listening, and speaking), language immersion (months spent in the FL country), and NL-FL similarity moderated the MFLE in personal vs. impersonal dilemmas. Results indicated a MFLE within personal dilemmas moderated by FL reading proficiency, whereby bilinguals with higher proficiency were less likely to make utilitarian decisions. The authors interpreted this finding by associating a higher reading proficiency in a FL to a purportedly stronger emotional intensity in processing FL materials, resulting in a larger proportion of emotion-driven (i.e., deontological) vs. cognitively controlled (i.e., utilitarian) decisions.

Overall, current findings suggest that not all bilinguals may experience a MFLE, and that multiple aspects of bilinguals' language background may differentially modulate response tendencies to moral conflict. The purpose of the present study is to shed additional explanatory light on the underlying mechanisms of the MFLE by investigating what dimensions of bilingual language experience contribute to distinct moral choices when moral dilemmas are presented in a FL. In particular, after testing the potential effect of FL on participants' decision-making, we evaluated what aspects of bilingualism – defined as a continuous and multidimensional phenomenon – affected bilinguals' choices. The use of a continuous approach to model the full spectrum of bilinguals' experiences and abilities has recently become popular in the cognitive neuroscience of bilingualism (e.g., Baum & Titone, 2014; De Luca et al., 2019; Sulpizio et al., 2020). A continuous approach has the advantage of (1) being more ecological, mirroring the intrinsic nature of bi-/multilingualism (i.e., heterogeneous across individuals and dynamic throughout life), and (2) allowing for a more precise modeling of the effects of bilingual experience. Here, we adopt this approach to investigate which bilingualism-related factors may influence the reasoning and judgment of bilingual speakers in the context of the MFLE.

1.1. (Putative) mechanisms responsible for the MFLE

It has been proposed that moral decisions result from the interplay of multiple processes that operate at different levels and time-scales (Greene & Haidt, 2002; Haidt, 2007; Moll et al., 2005). Greene's model of moral judgment (Greene et al., 2001, 2004) extended to moral cognition (and, particularly, to deontological/utilitarian assessments) the dual-processing architecture of theories common in the psychology of reasoning and decision-making (e.g., Kahneman & Frederick, 2002; Sloman, 1996). According to such theories, two types of separable and competing systems underpin human decisions: a fast, intuitive, automatic, nonvoluntary, and affect-driven system (System 1), and a slow, deliberate, systematic, voluntary, and cognitively controlled system (System 2). Greene's model associates System 1 with moral norms (i.e., deontological inclinations) and System 2 with dispassionate cost-benefit concerns (i.e., utilitarian inclinations). Decisions in harm-related scenarios would arise from the reconciliation of automatic emotional reactions to harm, relatively insensitive to the maximization of the aggregate welfare in a given scenario, and more deliberative cost-benefit appraisals that motivate harm acceptance to maximize outcomes. From a dual-process perspective, the larger proportion of utilitarian vs. deontological decisions or judgments associated with FL contexts could reflect either reduced emotional reactions to harm (i.e., the *reduced emotion*

hypothesis), or increased cognitive evaluations of outcomes (i.e., the *increased deliberation* hypothesis) (for a review, see Costa et al., 2019).

According to the *reduced emotion* hypothesis, actively thinking in a FL would lead to decisions or judgments that are less distorted by gut-feeling reactions (i.e., less deontological) because of the reduced emotional resonance associated with a foreign vs. a native language (Costa et al., 2014). A reduced emotional resonance in a FL would inhibit the automatic, affect-based response to a harmful action generated by System 1, leading to choices that are less affected by the emotional component of moral violations. This perspective is consistent with a large body of clinical, psychophysiological, and neuroimaging literature according to which a FL is generally perceived as less emotional than a NL (see Caldwell-Harris, 2015; Kazanas et al., 2019).

According to the *increased deliberation* hypothesis, actively thinking in a FL would lead to decisions or judgments that are more deliberative and cognitively controlled (i.e., more utilitarian) because of the higher cognitive load associated with processing a FL vs. a NL (Hayakawa et al., 2017; Keysar et al., 2012). The cognitive effort associated with text (or speech) comprehension in a FL, or the metacognitive effort associated with completing tasks in a FL, would enhance the cognitively controlled analysis of potential decision outcomes generated by System 2, leading to choices that are more concerned with the benefit-maximizing component of moral violations. This perspective is supported by evidence showing that an increased perception of disfluency or task difficulty in a FL may result in the activation of analytic processes that correct the output of more intuitive forms of judgment on logical reasoning tasks (e.g., Alter et al., 2007).

In the attempt to clarify whether the FLE primarily reflects reduced emotional reactions to harm (i.e., reduced reliance on System 1) or increased cognitive evaluations of outcomes (i.e., increased reliance on System 2), a process-dissociation procedure (Conway & Gawronski, 2013) was used that independently assesses deontological and utilitarian response inclinations in harm-related scenarios. The key feature of this procedure was the presentation of incongruent and congruent moral dilemmas. Traditional moral dilemmas are incongruent in the sense that causing harm violates deontology but upholds utilitarianism, and the sensitivity to norms and to aggregate consequences should lead to divergent responses. Congruent scenarios are structurally identical to incongruent ones, except that causing harm does not maximize aggregate outcomes. For example, if the choice concerns sacrificing one person to prevent five from being mildly injured (rather than killed), the harmful action is harder to justify on utilitarian grounds, but is equally unappealing from a deontological perspective. Therefore, the sensitivity to norms and to aggregate consequences should lead to the same response, although there remain nonmoral or antisocial reasons to accept harm. The process-dissociation procedure involves applying participants' responses to both incongruent and congruent dilemmas to a processing tree, and calculating two parameters for each participant which reflect the tendency to reject causing harm regardless of outcomes (i.e., deontological concerns), and the tendency to maximize outcomes regardless of whether doing so entails causing harm (i.e., utilitarian concerns) (Conway et al., 2018). By applying the process-dissociation approach to the FLE domain, Hayakawa et al. (2017) reported a reduction especially of deontological concerns elicited by FL use, whereas the effects on utilitarian concerns were weaker. Overall, FL contexts seemed to elicit dampened emotional reactions to harm rather than enhanced sensitivity to the consequences of

moral violations, thus aligning better with the *reduced emotion hypothesis* than the *increased deliberation hypothesis* (see Białek et al., 2019; Muda et al., 2018).

Broadly consistent with the idea that processing moral dilemmas in a FL reduces the sensitivity to moral norms is the *reduced access to norms hypothesis*, a further explanatory account of the FLE according to which a blunted deontology in FL settings is driven by a reduced accessibility of normative knowledge in FLs (Geipel et al., 2015a,b). There is a two-step argument behind this hypothesis. On the one hand, individuals are typically exposed to normative knowledge through social interactions mediated by their NL. On the other hand, autobiographical memories have been shown to include a trace of the language of encoding (e.g., Marian & Neisser, 2000; Schrauf & Rubin, 2000). On these grounds, a moral conflict presented in the NL would trigger greater language-dependent access to social and moral norms than a moral conflict presented in a FL, resulting in reduced deontological inclinations in FL settings (for a similar reasoning on socially relevant words, cf. Sulpizio et al., 2019).

1.2. (Putative) role of linguistic variables in modulating the MFLE

It is increasingly recognized that the malleability of moral choices in FL contexts may be conditional upon individual differences within dimensions of bilinguals' language experience. Interindividual variability along these dimensions is known to affect bilingual language processing (e.g., Fricke et al., 2019; Sulpizio et al., 2020), and may therefore underlie processing differences in decision-making when using a FL, with potential repercussions on the scope and characteristics of the MFLE. Research has begun to explore the putative influence of linguistic variables such as AoA, proficiency, and language dominance or use on bilinguals' patterns of moral responding, though with mixed or inconclusive results (see Del Maschio et al., 2022; Stankovic et al., 2022). Yet the role of these variables in modulating bilinguals' moral reasoning has been emphasized – albeit to varying degrees – by all previous theoretical accounts of the FLE, as detailed in the rest of this section.

FL AoA refers to the age when FL learning begins. A role of FL AoA has been posited by the *reduced emotion hypothesis*, the *increased deliberation hypothesis*, and the *reduced access to norms hypothesis*. Under these accounts, a later acquired language as compared to one acquired at an early age is expected to promote a larger proportion of utilitarian vs. deontological decisions or judgments. This effect can be attributed to a weaker emotional attachment in the later acquired language, as predicted by the *reduced emotion hypothesis*, or to an increased cognitive effort associated with processing a language acquired beyond critical or sensitive periods in development, as implied by the *increased deliberation hypothesis*. Alternatively, a sharpened utilitarianism or a blunted deontology could stem from a reduced language-dependent activation of moral norms to which individuals are usually first exposed in childhood, as predicted by the *reduced access to norms hypothesis*.

FL proficiency refers to the level of competence a bilingual has in the production and understanding of a FL. A role of FL proficiency has been posited by the *reduced emotion hypothesis* and the *increased deliberation hypothesis*. Under both accounts, a lower vs. a higher level of proficiency in a FL is expected to promote a larger proportion of utilitarian vs. deontological decisions or judgments. This may be due to a reduced emotional impact associated with a limited language proficiency in the FL,

as the *reduced emotion* hypothesis would suggest, or because of an increased cognitive effort associated with FL processing in poorly proficient bilinguals, as predicted by the *increased deliberation* hypothesis.

Language dominance is a complex construct that includes “a linguistic proficiency component, an external component (input), and a functional component (context and use)” (Montrul, 2016, p. 16). The variable is critical in characterizing asymmetries of skill or use of a bilingual’s language over the other, and in discriminating between different types of bilingual speakers such as balanced and dominant bilinguals. Although the number of studies that have specifically addressed the contribution of language dominance to the MFLE is scarce, there is some evidence indicating that this factor modulates moral decisions in FL settings (e.g., Wong & Ng, 2018). A FL that is weaker than the NL might promote a larger proportion of utilitarian vs. deontological decisions or judgments in the FL. On the other hand, more balanced types of bilingualism could decrease utilitarian tendencies in the FL, leveling out differences in decision processing between languages. According to the *reduced emotion* hypothesis, a heightened utilitarianism or a blunted deontology would depend on the different emotional resonance across a bilingual’s languages, possibly modulated by the relative proficiency in or exposure to the FL vs. the NL. Alternatively, as implied by the *increased deliberation* hypothesis, a more marked utilitarianism in FL settings may derive from the effort associated with processing a FL that is used less proficiently or frequently than the NL.

1.3. The present study

Given the nature of bilingualism as a construct comprising several interrelated dimensions, it is overall unlikely that a single, segregated feature of bilinguals’ language experience may be responsible for divergent patterns of moral responses in FL contexts. However, prior literature on the MFLE has mainly studied bilingualism from a narrowly defined perspective, characterizing this experience as a qualitative rather than a quantitative phenomenon, or focusing on a single linguistic variable and not others. Our study aims to bridge this gap by adopting a more articulated perspective that takes into account the extent to which individuals vary as bilinguals along multiple dimensions. A similar approach can help us identify which linguistic factors modulate bilinguals’ decision-making and judgment in high-conflict situations. Crucially, understanding for whom and in what circumstances decision-making differences occur may also inform previous alternative accounts of the FLE.

First, in a between-group analysis, we investigated in three sacrificial dilemmas (*Surgeon, Factory, Bike week*) the effect of language context on the moral inclinations of native Italian speakers assigned to either a NL condition (i.e., Italian) or a FL condition (i.e., English). We also tested whether response tendencies were qualified by (1) the perceived permissibility of the moral violation depicted in each scenario, and (2) the perceived emotional distress associated with processing each scenario. (1) is motivated by explanations relating the MFLE to the reduced mental accessibility of normative knowledge in FLs. Moreover, it is also motivated by data showing that the framing of moral questions – tapping either willingness to commit a harmful action, or judgment about whether the action is permissible – differently affects moral reasoning and decision outcomes. (2) is motivated by explanations relating the MFLE

to reduced emotional responses in FLs. (1) and (2) may also help discriminate between norm- and emotion-based accounts of the MFLE.

Second, we focused on the participants assigned to the FL condition and evaluated whether their moral inclinations, together with their norm judgments and distress ratings, were modulated by the individual and interactive effects of the main quantifiable dimensions of bilingual language experience (i.e., FL AoA, FL proficiency, and Language dominance).

2. Materials and methods

2.1. Participants

One hundred ninety-eight native Italian speakers (133 F; $M_{\text{age}} = 24.89$ years, $SD = 3.99$) volunteered to participate in the study. All participants were young adults who were born and raised in Italy. Educational history (in years) was collected for all participants ($M = 15.99$, $SD = 2.11$) by means of a written questionnaire that asked participants to report how many years of formal education they had completed. Ninety-six participants were randomly assigned to the native language condition (NL; Italian), one-hundred and two ($N = 102$) to the foreign language condition (FL; English). All participants in the FL condition spoke English as a second language (L2). No significant differences were found between participants assigned to the NL and the FL condition for the matching criteria of age (NL: $M = 24.3$, $SD = 3.67$; FL: $M = 25.4$, $SD = 4.25$; $W = 5,391$, $p = 0.076$), gender (NL = 61 F, 35 M; FL: 71 F, 27 M; $\chi^2 = 1.3835$, $p = 0.23$) and education (NL: $M = 15.7$, $SD = 2.13$; FL: $M = 16.2$, $SD = 2.08$; $W = 5,412$, $p = 0.061$).

The study was conducted with ethical approval from the Human Research Ethics Committee of the San Raffaele Hospital (Milan, Italy). Informed consent was obtained from all participants.

2.2. Bilingual language background

The bilingual language background of participants in the FL condition was assessed by means of self-report and objective measures. The Language History Questionnaire (version 3) (LHQ3) (Li et al., 2020) was used to assess participants' FL AoA, FL proficiency, and language dominance. FL AoA is operationalized as the lowest age at which participants begin to listen to or learn to speak or write in an FL (i.e., English). FL proficiency is computed as the weighted sum of participants' self-rated proficiency levels on different components of the FL (i.e., listening, speaking, reading, and writing). For each language, participants had to self-rate their abilities on a seven-point Likert scale ('Very poor'; 'Poor'; 'Limited'; 'Average'; 'Good'; 'Very Good'; 'Excellent'). Daily usage of FL is estimated as daily hours spent using the FL in different activities and conversations. The language dominance score is an aggregate score of both the participants' self-reported proficiency in a language and the estimated time spent every day using that language. Language dominance is ultimately computed as the relative ratio of the dominance score of each language (i.e., FL = English, NL = Italian) against that of the native language (i.e., Italian). The ratio score ranges from 0 to 1 and can be used to determine whether a participant is equally exposed to both FL and NL, or whether one language is dominant over the

other (0 = the participant is exposed only to NL; 1 = the participant is equally exposed to NL and FL).

An English Proficiency Test developed by Transparent Language (freely available online at <https://www.transparent.com/>) was used as an objective measure of FL proficiency. The test consists of a 40-item questionnaire divided into multiple sections, including 30 questions evaluating the knowledge of English grammar (e.g., participants had to choose the best grammar option to fill in the blanks within English sentences), and 10 questions evaluating text comprehension ability (i.e., participants had to answer questions referred to four short English texts) (EPT scores range: 0–40). Furthermore, FL vocabulary knowledge was assessed by asking participants to name 30 pictures selected from the Snodgrass and Vanderwart (1980)'s dataset (Snodgrass scores range: 0–30). The frequency per million of each corresponding word was extracted, and then 10 low-frequency, 10 medium-frequency, and 10 high-frequency words were identified, setting the 30th and 60th percentile, calculated over the pictures' dataset, as cut-offs.

The descriptive statistics for the language background measures of participants assigned to the FL condition are reported in Table 1.

2.3. Materials

Three moral dilemmas (*Surgeon*, *Factory*, *Bike week*; adapted from Cecchetto et al., 2017), together with two filler dilemmas (*Train or bus*, *Plant transport*; Geipel et al., 2015a) were used as stimuli (see Supplementary Materials for the full text). *Surgeon*, *Factory*, and *Bike week* are sacrificial dilemmas in which the decision-maker has to choose between killing one person to save a group of people (utilitarian judgment), or not killing one person and letting the group die (deontological judgment). Based on Christensen et al. (2014), we classified our moral dilemmas in terms of *Personal Force*, *Benefit Recipient*, *Evitability*, and *Intentionality*, and held these factors constant across the three scenarios. All moral dilemmas were classified as personal, other-beneficial, avoidable, and instrumental. A dilemma is personal (vs. impersonal) when it requires up-close and personal action that leads to harm to a specific individual or group. In particular, a dilemma is personal when the force that directly impacts the other person(s) is generated by the decision maker's muscles, or when the decision-maker pushes another one with one's hands or with a rigid object. A dilemma is other-beneficial (vs. self-beneficial) when the decision maker's life is not at risk, and avoidable (vs. inevitable) when the sacrificed life would be spared as a

Table 1. Linguistic background of participants assigned to the foreign language condition (English)

	<i>M</i> ± <i>SD</i> (range)
FL AoA (LHQ3)	7.89 ± 3.35 (0–12)
FL proficiency (LHQ3)	0.74 ± 0.11 (0.43–1)
FL daily usage (LHQ3)	4.52 ± 4.01 (0–18)
Language dominance (LHQ3)	0.65 ± 0.11 (0.36–0.95)
FL proficiency (EPT)	34.95 ± 3.22 (21–40)
FL vocabulary (Snodgrass)	20.81 ± 5.12 (5–30)

Means (*M*), standard deviations (*SD*) and range are reported for FL AoA, FL Proficiency, and language dominance. Scores for FL proficiency (LHQ3) and language dominance (LHQ3) range from 0 to 1; scores for FL proficiency (EPT) range from 0 to 40; scores for FL vocabulary (Snodgrass) range from 0 to 30. LHQ3 = Language History Questionnaire (version 3); EPT = English Proficiency Test.

result of an alternative course of action. A dilemma is instrumental (vs. incidental) when sacrificing one life is not an unintended consequence of a given course of action (i.e., a side-effect), but rather a deliberate harmful act intended to save a greater number of people (see Christensen et al., 2014). To rule out possible biases produced by in/out group differences (e.g., Swann et al., 2010; Uhlmann et al., 2009), the age, gender, socioeconomic status, ethnicity, and cultural identity of the individuals described in each scenario were not specified. Moreover, consistent with the evidence that the number of individuals saved as a result of a moral violation affects the degree of utilitarianism in decision-making (e.g., Cao et al., 2017), the number of saved lives in our three scenarios ranged from five to seven, following the categorical distinction proposed by Christensen et al. (2014) (1, 5–10; 2, 11–50; 3, 100–150; 4, ‘thousands’ or ‘masses’ of people). Since also the structure (i.e., the presentation order of relevant information), the expression style (e.g., the amount of descriptive and dramatic language), and the wording of a moral dilemma can bias moral judgments (e.g., Borg et al., 2006; O’Hara et al., 2010), we kept these factors as homogeneous as possible for our three scenarios and their language versions. The word count of the dilemmas in both languages was also kept as equal as possible. The original materials were written in English, and then translated into Italian by a proficient bilingual. Two independent judges controlled the translated versions for consistency with the English version.

Differently from *Surgeon*, *Factory* and *Bike week*, *Train or bus* and *Plant transport* are filler dilemmas that do not tap deontological vs. utilitarian inclinations, but rather choices between self-interest and common good in nonsacrificial contexts. *Train or bus* involves a decision between traveling by bus vs. train given certain time constraints; *Plant transport* involves a decision between doing multiple (vs. a single) car trips to avoid ruining a car’s upholstery. In the *Train or bus* dilemma, a *yes* response (traveling by train to attend a meeting on time) was expected to induce a high rate of endorsements. In the *Plant transport* dilemma, a *no* answer (doing a single car trip to avoid polluting) was expected to induce a high rate of endorsements (see Brouwer, 2019; Geipel et al., 2015b). If there would be language misunderstandings, the rate of *yes/no* responses should be around 50%. The word count of the English and Italian versions of the filler dilemmas was kept as equal as possible. The original materials were written in English, and then translated into Italian by a proficient bilingual. Two independent judges controlled the translated versions for consistency with the English version. The percentage of *yes* responses for *Train or bus* was 96.93% for the FL condition and 96.87% for the NL condition. On the other hand, a substantial number of participants reported – during data collection – that they did not know the meaning of a word (*upholstery*) that was key to the understanding of the *Plant transport* scenario. To avoid biases related to poor comprehension, we did not analyze the percentage of responses for this filler dilemma.

2.4. Procedure

The experiment was implemented using Google Forms and presented in the lab on a computer screen to all participants. No manipulation of time pressure was exerted to trigger participants’ responses. Background and linguistic measures were collected, for each participant, after the presentation of moral dilemmas. The whole session lasted ~15 minutes for participants assigned to the NL condition, and ~45 minutes for participants assigned to the FL condition. The difference in the duration of the

experimental session was due to the additional measures collected for the FL group. To ensure that participants were fully informed about the study and that there were no misunderstandings attributable to low FL proficiency, information about the study and informed consent were presented, for all participants regardless of experimental condition, in the participants' NL (Italian).

The presentation order of moral and filler dilemmas was randomized across participants. Dilemmas were followed by multiple questions presented in a fixed order for all participants. Following each moral dilemma, a dichotomous question was presented to all participants asking whether they would choose to carry out the depicted moral violation (*Do you decide to... yes = utilitarian response; no = deontological response*). All participants were then required to evaluate the permissibility of the moral violation on a 7-point scale (*How do you rate this action on a scale from 1 to 7? 1 = forbidden, 4 = permissible, 7 = obligatory*). The first (binary) question required participants to think about whether they would cause harm to maximize overall outcomes. The second (7-point scale) question required participants to judge the permissibility of the moral violation prospected in the scenario. After the moral response and the norm judgment questions, all participants had to rate the extent to which the dilemma made them feel distressed on three 7-point scales, each referring to a particular negative emotion or state of being (*Thinking about the scenario I just read, I felt...upset, worried, sad. 1 = not at all, 4 = somewhat, 7 = very much*; from Geipel et al., 2015a). In addition, participants assigned to the FL condition were asked to indicate to what extent they understood the dilemma on a 7-point scale (*Did you understand the English text in which the problem was presented? 1 = not at all, 4 = average, 7 = very well*).

Following each filler dilemma, a dichotomous question was presented to all participants asking whether they would choose to carry out the depicted action (*Do you decide to... yes = self-interest response; no = common good response*). In addition, participants assigned to the FL condition were asked to indicate to what extent they understood the dilemma on a 7-point scale (*Did you understand the English text in which the problem was presented? 1 = not at all, 4 = average, 7 = very well*). Participants who rated their understanding of at least one moral dilemma as two or less were set to be excluded from subsequent analyses. Four participants were excluded on the basis of this criterion, leading to a final sample of 98 ($N = 98$) participants assigned to the FL condition and 96 ($N = 96$) participants assigned to the NL condition.

2.5. Statistical analyses

All analyses were performed using R (R Core Team, 2015). The impact of FL on participants' choices was assessed by means of two complementary analytic approaches.

First, in a between-groups analysis, we evaluated whether the language context (i.e., NL vs. FL) affected participants' moral decisions, norm judgment, and emotional distress. This approach follows the one typically used in studies investigating the MFLE, in which different groups of participants speaking an NL and an FL are compared. Second, in a within-group analysis involving participants exposed to the FL context only, we evaluated what dimension(s) of bilingual experience (i.e., FL

AoA, FL proficiency, and Language dominance) affected the outcome variables aforementioned.

2.5.1. Between-groups analysis (evaluating the effect of language context)

To evaluate the impact of FL on moral decisions, we ran, for each moral dilemma, a logistic regression model with type of response (*yes/no*) as a dependent variable and language context (NL vs. FL) as a predictor.

To evaluate the impact of FL on norm judgment, we ran, for each dilemma, an ordinal logistic regression with action permissibility ratings as the dependent variable and language context (NL vs. FL) as the predictor. The model was run using the MASS library (Venables & Ripley, 2013).

With regard to the perceived emotional distress, for each moral dilemma, since the three distress items (*upset*, *worried*, and *sad*) were highly correlated (Cronbach's alpha was 0.87 for *Surgeon*, 0.89 for *Factory*, and 0.91 for *Bike week*), we collapsed them into a single index, and computed the mean over the three scales. To evaluate the impact of FL on perceived emotional distress, we ran, for each dilemma, a linear regression with the mean distress rating as dependent variable and language context (NL vs. FL) as predictor.

2.5.2. Within-group analysis (evaluating the effect of bilingual experience)

Focusing on participants in the FL condition, we evaluated, for each dilemma, whether (and to what extent) the different dimensions of bilingual experience affected moral decisions, norm judgment, and emotional distress. To avoid multicollinearity among predictors, we checked, before the analyses, for any strong correlations ($r > 0.50$, see e.g., Taylor, 1990) among the measures of bilingual language experience. Supplementary Figure S1 reports the results of Spearman's correlations. FL self-reported proficiency (LHQ3) was highly correlated with language dominance ($\rho > 0.8$), and vocabulary knowledge (Snodgrass) was moderately correlated with FL Proficiency (EPT) ($\rho = 0.35$). FL Proficiency (EPT) was entered into the statistical models as an objective measure of FL competence that assesses both grammatical and conversational knowledge and text comprehension. Vocabulary knowledge (Snodgrass) and FL self-reported proficiency (LHQ3) were excluded from subsequent analyses. Overall, based on the correlations' results, the following measures were entered as predictors in the statistical models: FL AoA, FL proficiency (EPT), and Language dominance. Because of the large heterogeneity in their scales, all predictors were centered, before being entered into the models, by subtracting, for each predictor, its mean to each value.

To evaluate the impact of bilingual experience on moral decisions, we ran, for each dilemma, a logistic regression model with type of response (*yes/no*) as the dependent variable and FL AoA, FL proficiency, and Language dominance as predictors. Effects were assessed via likelihood ratio tests comparing models in which the effect under examination was present vs. absent. Terms were retained only in case their exclusion would determine a significant decrease in goodness of fit. In case any interaction resulted significant, all the lower-order terms involved were retained.

To evaluate the impact of bilingual experience on norm judgment, we ran, for each dilemma, an ordinal logistic regression with the permissibility ratings as the dependent variable and FL AoA, FL proficiency, and Language dominance as predictors. The main effects and their interactions were evaluated as described above.

Finally, to evaluate the impact of bilingual experience on perceived emotional distress, we ran, for each dilemma, a linear regression with the mean distress rating (computed as the mean of the ratings given to the three distress questions) as dependent variable and FL AoA, FL proficiency, and Language dominance as predictors. The main effects and their interactions were evaluated as described above.

3. Results

3.1. Between-group analysis (evaluating the effect of language context)

The percentage of *yes* (utilitarian) responses, the ratings for the perceived permissibility of moral violations, and the ratings for the perceived emotional distress are reported, for each moral dilemma, in Fig. 1.

A significant effect of language context on moral decisions was observed for the *Bike Week* dilemma ($\chi^2(1) = 4.62, p = 0.03, \text{beta} = 0.65, \text{std. err.} = 0.30, z = 2.13$), with the use of an FL increasing the rate of utilitarian responses. In particular, while 28.12% of participants decided to push the biker off road when the dilemma was presented in the NL (i.e., Italian), this rate increased to 42.85% when it was presented in an FL (i.e., English). No significant effect of language context was observed for the *Surgeon* ($\chi^2 = 2, p > 0.1$) and the *Factory* dilemmas ($\chi^2 < 1, p > 0.4$).

Moreover, a significant effect of language context was observed on the perceived permissibility of the moral violation prospected in both the *Surgeon* ($\text{beta} = 0.36, \text{std. err.} = 0.12, t = 2.42, p = 0.015$) and the *Bike week* ($\text{beta} = 0.36, \text{std. err.} = 0.13, t = 2.82, p = 0.004$) dilemmas, with higher permissibility ratings associated with the use of an FL compared to the NL. No significant effect of language context was observed on the perceived permissibility of the moral violation prospected in the *Factory* dilemma ($t < 1, p > 0.8$).

With respect to the impact of language context on the perceived emotional distress associated with processing each scenario, we did not find any significant difference between the FL and the NL conditions (all $ps > 0.2$).

3.2. Within-group analysis (evaluating the effect of bilingual experience)

In what follows, for clarity, we separately report the results of each dependent variable (the parameters of all statistical models are reported in Supplementary Tables S1–S9).

3.2.1 Moral decisions

In the *Surgeon* dilemma, the three-way interaction between FL proficiency, Language Dominance, and FL AoA was significant ($\chi^2(1) = 8.296, p = 0.003, \text{beta} = 1.43, \text{std. err.} = 0.68, z = 2.08$). To better understand this result, we inspected the interaction by splitting the data into early (i.e., FL AoA $\leq 6, N = 39$) and late bilinguals (i.e., FL AoA $> 6, N = 59$). As shown in Fig. 2, early and late bilinguals display (almost) opposite patterns of moral behavior. In early bilinguals, the probability of utilitarian responses increased with the increase of FL proficiency when the dominance ratio was low, whilst the same probability decreased with the decrease of FL proficiency when the dominance ratio was high. In late bilinguals, the probability of utilitarian decisions decreased with the decrease of FL proficiency when the dominance ratio

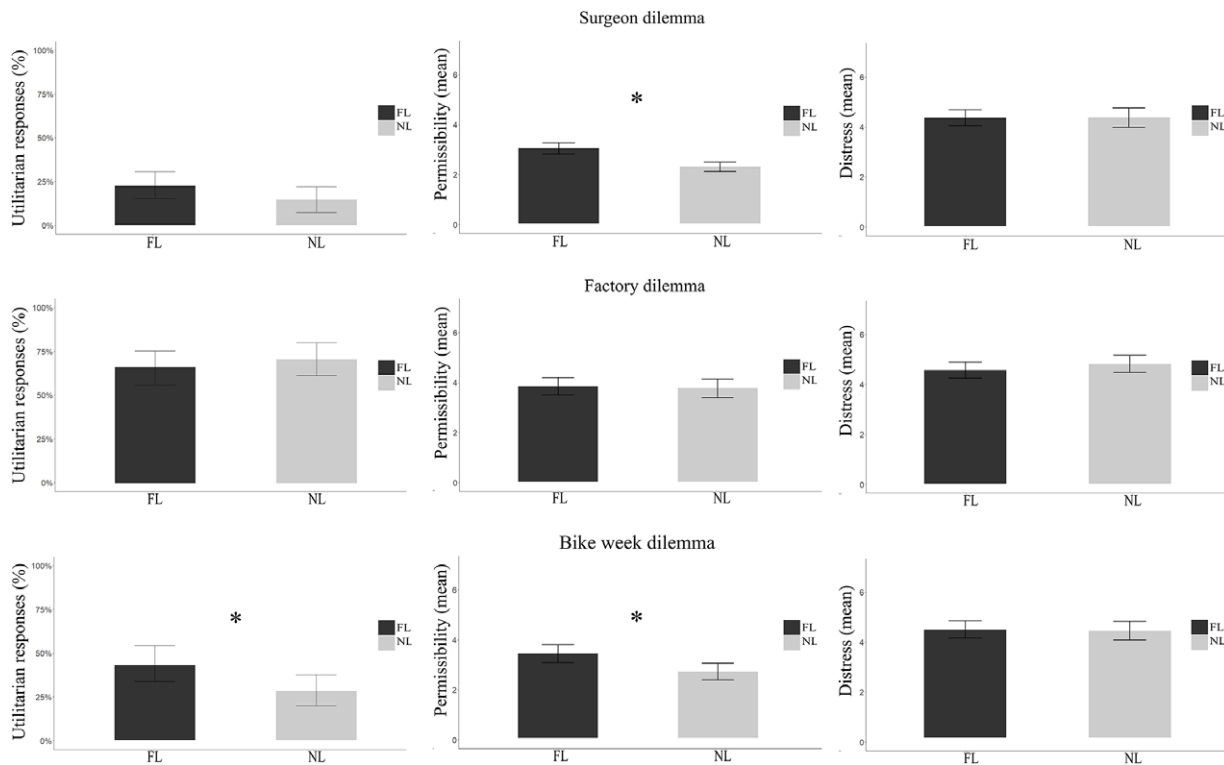


Fig. 1. The percentage of utilitarian responses (left panels), the mean permissibility of the moral violation (central panels), and the mean emotional distress (right panels) are reported for each scenario (A = Surgeon, B = Factory, C = Bike week) and for language condition (FL = foreign language - $N = 98$; NL = native language - $N = 96$). Significant differences between language conditions are marked with an asterisk ($p < 0.05$).[†]

[†]This figure was originally published with an error in the participant responses. This has now been corrected and a corrigendum is available at: <https://doi.org/10.1017/langcog.2022.32>

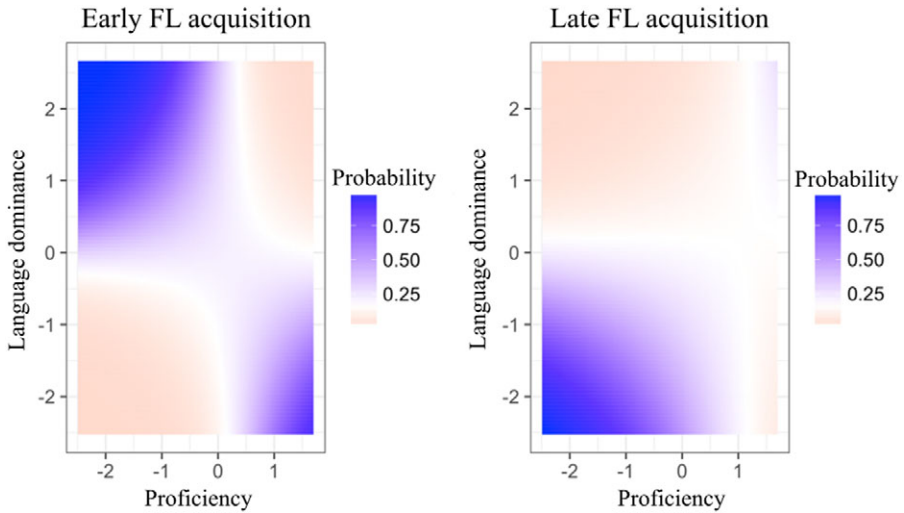


Fig. 2. Effects of bilinguals' ($N = 98$) language experience on moral decisions in the *Surgeon* dilemma. Panel (A) and (B) show the language dominance \times FL proficiency interaction for early and late bilinguals, respectively. Bilinguals within the 1st and the 4th quartile were selected to represent early and late bilinguals. The values represent the probability predicted by the statistical model. The axes display centered predictors.

was low, whilst utilitarian tendencies appeared constantly low when the dominance ratio was high.

No significant effect of bilingual experience on the probability of utilitarian decisions emerged for the *Factory* and the *Bike Week* dilemmas.

3.2.2 Perceived permissibility of moral violations

The analysis of the *Surgeon* dilemma revealed a significant effect of FL Proficiency \times FL AoA interaction ($\chi^2(1) = 4.288$, $p = 0.03$, $\beta = -0.45$, $\text{std. err.} = 0.21$, $t = -2.06$). We inspected the interaction by separately looking at early (i.e., FL AoA ≤ 6 , $N = 39$) and late bilinguals (i.e., FL AoA > 6 , $N = 59$). As shown in Fig. 3, for early bilinguals, the probability to judge the action as forbidden decreases with the increase in FL proficiency, whereas the other judgments tend to be almost constant. For late bilinguals, an increase in proficiency mainly affects extreme judgments by increasing the probability to judge the action as forbidden and decreasing the probability to judge the action as obligatory, which both reach levels similar to those of early, high proficient bilinguals.

The analysis of the *Bike Week* dilemma showed two significant interactions between FL AoA and Language Dominance ($\chi^2(1) = 4.383$, $p = 0.03$, $\beta = -0.47$, $\text{std. err.} = 0.22$, $t = -2.07$), and between FL Proficiency and Language Dominance ($\chi^2(1) = 3.95$, $p = 0.04$, $\beta = -0.30$, $\text{std. err.} = 0.15$, $t = -1.95$). The first interaction was inspected by separately looking at early (i.e., FL AoA ≤ 6 , $N = 39$) and late bilinguals (i.e., FL AoA > 6 , $N = 59$). As shown in Fig. 3, in early bilinguals, when the dominance ratio was high, the action was less likely to be judged as forbidden, and slightly more likely to be acceptable. In late bilinguals, the pattern was the same, but slightly more marked. The interaction between FL Proficiency and Language

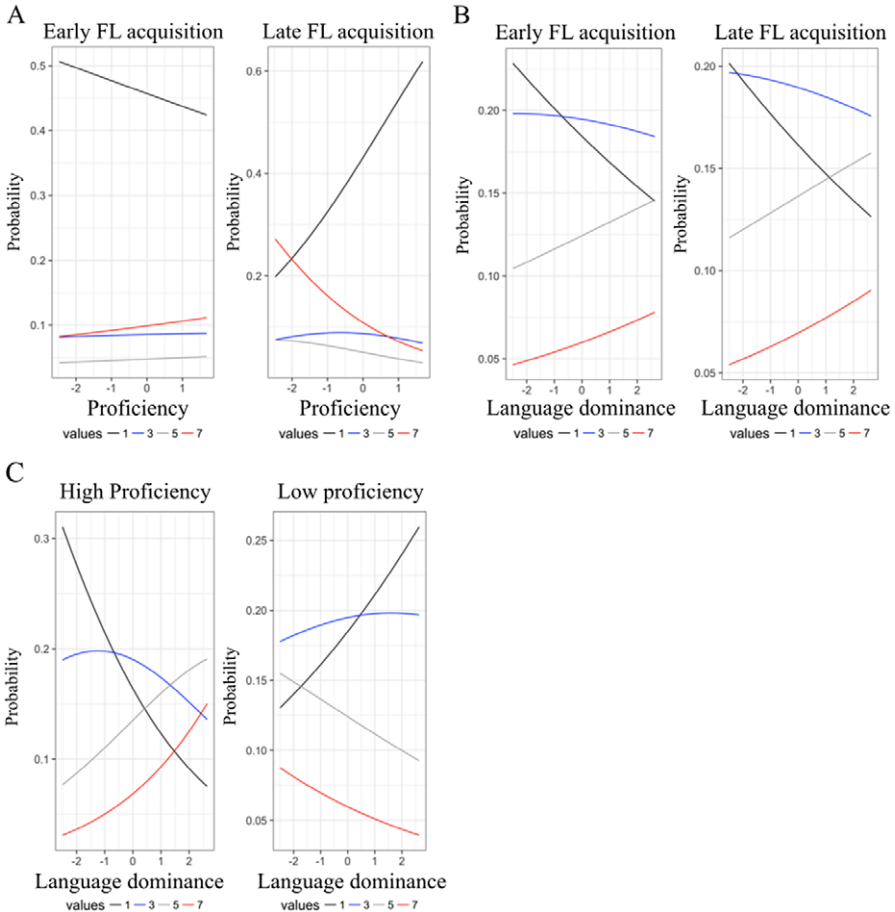


Fig. 3. Effects of bilinguals' ($N = 98$) language experience on the perceived permissibility of moral violations. (A) FL AoA \times FL Proficiency interaction in the *Surgeon* dilemma. (B) FL AoA \times Language Dominance interaction in the *Bike week* dilemma. (C) FL Proficiency \times Language Dominance interaction in the *Bike week* dilemma. Bilinguals within the 1st and the 4th quantile of FL AoA were selected to represent early and late bilinguals, respectively. Bilinguals within the 1st and the 4th quantile of FL Proficiency were selected to represent low and high proficient bilinguals, respectively. The values represent the probability predicted by the statistical model. Different lines represent predicted probabilities for different representative values of the permissibility scale. The x-axis displays the centered predictor.

Dominance was inspected by separately looking at low and high proficient bilinguals (considering the 1st and the 4th quantile of FL Proficiency as cut-off points, respectively). In highly proficient bilinguals, when dominance was high, the action was less likely to be judged as forbidden, and more likely to be acceptable. Low proficient bilinguals, instead, showed an opposite pattern: higher Language Dominance was associated with a higher probability to judge the action as forbidden and a lower probability to judge the action as acceptable/permissible.

No significant effect of bilingual experience on the perceived permissibility of moral violations emerged in the *Factory* dilemma.

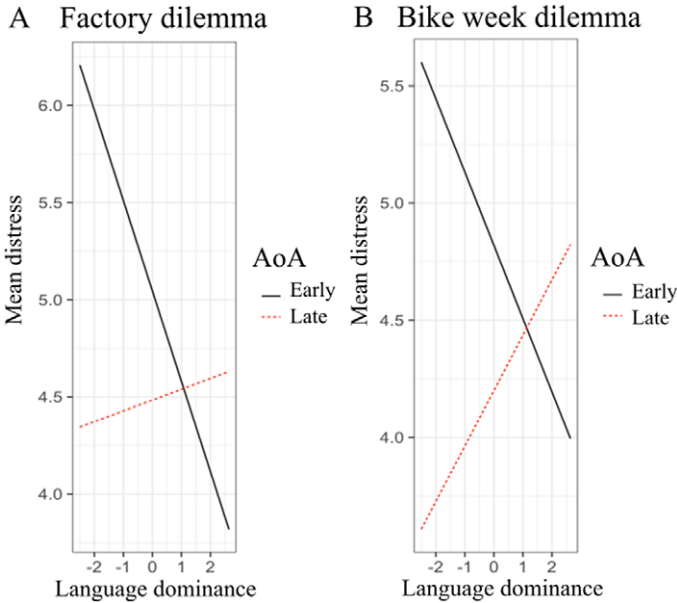


Fig. 4. Perceived emotional distress as a function of bilinguals' ($N = 98$) language experience and dilemma type. (A) Effects of FL AoA and Language Dominance on perceived distress in the *Factory* dilemma. (B) Effects of FL AoA and Dominance ratio on perceived distress in the *Bike week* dilemma. Bilinguals within the first and the fourth quantile were selected to represent early and late bilinguals, respectively. Lines represent values predicted by the statistical model. The x -axis displays the centered predictor.

3.2.3 Perceived emotional distress

The analysis of the *Factory* dilemma showed a significant interaction between FL AoA and Language Dominance ($\chi^2(1) = 97.377, p = 0.005, \text{beta} = 0.44, \text{std. err.} = 0.15, t = 2.77$). While in late bilinguals the higher the dominance the higher the distress perception, in early bilinguals the pattern was the opposite and more marked (see Fig. 4). Also, the main effect of FL Proficiency was significant ($\chi^2(1) = 96.273, p = 0.009, \text{beta} = 0.46, \text{std. err.} = 0.17, t = 2.56$), indicating that the higher the FL proficiency, the more distressful the dilemma was perceived.

The analysis of the *Bike week* dilemma also showed a significant interaction between FL AoA and Language Dominance ($\chi^2(1) = 108.74, p = 0.005, \text{beta} = 0.46, \text{std. err.} = 0.16, t = 2.75$). While in late bilinguals the higher the dominance ratio the higher the distress perception, in early bilinguals the pattern was the opposite (Fig. 4).

No significant effect of bilingual experience on perceived emotional distress emerged for the *Surgeon* dilemma.

4. Discussion

It is still uncertain as to what aspects of bilinguals' language experience contribute to the MFLE. To bridge this gap, we extended to the FLE domain a conceptualization of bilingualism that frames the phenomenon as a composite and continuous measure, and explored the contribution of FL AoA, FL proficiency, and language dominance in modulating bilinguals' responses to moral conflict in harm-related scenarios. In the

next paragraphs, we first discuss participants' ($N = 196$) differences in moral decisions and judgment as a function of language context (i.e., the 'classic' MFLE); then, we discuss whether (and to what extent) moral decision-making in a FL is influenced by variability in the language experiences of bilingual speakers ($N = 98$).

4.1. Evaluating the effect of language context

A significant effect of language context on the participants' probability to cause outcome-maximizing harm was observed in the *Bike week* dilemma, but not in the *Surgeon* and the *Factory* dilemmas. Null findings are not surprising *per se*, as other studies also failed to obtain an MFLE in harm-related scenarios (e.g., Białek et al., 2019; Brouwer, 2019 – Experiment 1; Chan et al., 2016; Muda et al., 2018). Somewhat more surprising, albeit not utterly unexpected, is the fact that the MFLE was constrained to a single dilemma. Indeed, the scenarios we used as experimental stimuli were standardized in terms of design parameters (all were classified as personal, other-beneficial, avoidable, and instrumental), and kept as homogeneous as possible in terms of information structure, expression style, and wording (both within and across languages). The lack of evidence for an MFLE generalizing to dilemmas other than *Bike week* may indicate that there are unique content features in some dilemmas that cue more utilitarian or less deontological responding. One might speculate that the peculiarity of these features lies in their emotional connotation (see, e.g., Carmona-Perera et al., 2013; Ugazio et al., 2012). If some scenarios are perceived as more emotionally arousing than others because of certain features, the emotion-reducing effect of the FL may be selectively stronger for those scenarios, narrowing down the scope of the MFLE. However, when we tested the immediate (negative) perceived emotions that participants' felt when processing our set of stimuli, no significant difference emerged between the FL and the NL condition, nor between scenarios in the FL condition. This finding not only seems to disconfirm the hypothesis aforementioned, but also fails to corroborate the claim that the MFLE is driven or mediated by reduced emotionality in FL settings (for previous studies reporting a lack of connection between emotions and the MFLE, e.g., Chan et al., 2016; Geipel et al., 2015a,b; Miozzo et al., 2020; Muda et al., 2020). We tentatively hypothesize that the contextualization of each scenario might play a role in its consequent perception, apparently through processing routes that are not based on affect heuristics, but rather on normative principles triggered by context-specific cues. Partial support to this interpretation comes from norm judgment ratings. Indeed, the MFLE observed for *Bike week* was qualified by a significant effect of language context on the perceived permissibility of the violation prospected in that scenario. In particular, higher permissibility ratings were observed for participants in the FL vs. NL condition. Coupled with a higher willingness to cause outcome-maximizing harm, this finding seems broadly consistent with the *reduced access to norms* hypothesis, according to which a FL promotes less condemnation of violations of social and moral norms than the NL (e.g., Gawinkowska et al., 2013; Geipel et al., 2015a,b). Interestingly, compared to participants in the NL condition, participants who used a FL rated as more permissible also the moral violation prospected in the *Surgeon* dilemma, in the face of a null effect of language context on outcome-maximization tendencies. These data may hint at dissociable mechanisms underlying first-person predictions (elicited by questions such as 'Do you decide to perform the

described action?') and judgments about the permissibility of moral transgressions (elicited by questions such as 'How do you rate the described action on a scale from forbidden to obligatory?'). Whether a particular framing of moral questions differently affects decision-making outcomes is currently debated among moral psychologists (see, for a critical review, Malle, 2021). However, evidence from both behavioral and brain imaging studies suggests that questions eliciting first-person predictions do not seem to trigger the same processing as norm judgments (Garrigan et al., 2016), and do not lead to the same patterns of results (Tassy et al., 2013). Our findings support the conceptual division between moral decisions and norm judgments also in FL settings, suggesting that the framing of moral questions might be critical to elicit or not an MFLE.

4.2. Evaluating the effect of bilingual language experience

A joint effect of FL AoA, FL proficiency, and language dominance on the bilinguals' probability to cause outcome-maximizing harm was observed in the *Surgeon* dilemma, whereas no main or interactive effects emerged for the *Factory* and the *Bike week* dilemmas. Recall that previous results from the between-group analysis on the same outcome variable revealed a significant difference between language conditions for the *Bike week* dilemma, but not for the *Surgeon* and the *Factory* dilemmas. As a whole, two inferences can be drawn from this pattern of findings. On the one hand, it appears that bilinguals' language background can influence decision-making in a FL, without this necessarily emerging into a language of presentation effect. On the other hand, a significant effect of language context on decision tendencies, when detected, may depend on factors other than differences in the language experiences of bilingual speakers.

The three-way interaction between FL AoA, FL proficiency, and language dominance in the *Surgeon* dilemma was inspected by separately looking at early and late bilinguals, who showed divergent decision patterns. In particular, early bilinguals showed increased utilitarianism with the increase of FL proficiency when the NL was dominant, whilst utilitarian tendencies decreased with the decrease of FL proficiency in the presence of a more balanced exposure to both languages. By contrast, late bilinguals showed decreased utilitarianism with the increase of FL proficiency when the NL was dominant, whilst utilitarian tendencies appeared constantly low in the presence of a more balanced exposure to both languages. According to both the *reduced emotion* hypothesis and the *increased deliberation* hypothesis, higher levels of proficiency in and exposure to the FL would decrease utilitarian responding in FL settings. Our findings outline a more complex scenario, not fully compatible with this prediction. It is worth remarking, however, that the major explanatory hypotheses of the FLE are predominantly based on data from a specific group of bilinguals (often termed 'FL learners') – that is, unbalanced speakers who have learned their FL in classroom settings, usually in countries where their NL was dominant. In such cases, the increase in utilitarian decisions when using a FL was typically larger for low than for high proficiency participants (e.g., Corey et al., 2017; Costa et al., 2014; Geipel et al., 2015b; but see Mills & Nicoladis, 2020). Our findings from late bilinguals are broadly consistent with this previous evidence, but also point to language dominance as a key factor that may distinctively shape moral reasoning in a late acquired FL. More in general, the current results – which must be replicated with larger

samples – indicate that individual differences in the relationship between proficiency and dominance differentially affect decision-making in an FL as a function of AoA or learning context. Recent neuroimaging accounts have shown that proficiency, usage or dominance, and AoA interact to shape both the functional and structural organization of the bilingual brain (e.g., Fedeli et al., 2021; Gullifer et al., 2018; Sulpizio et al., 2020). The present findings suggest that variability along these interrelated dimensions also modulate processing differences in decision-making when using a FL, further endorsing a continuous and multicomponential approach to bilingualism as a life experience. Importantly, this specific pattern of results also highlights that some variables related to bilinguals' language experience, albeit conceptually distinct, may be partially overlapping, as in the case of AoA and learning context. Indeed, the early bilinguals in our sample were also those who learned their FL in nonformal settings, whereas the late bilinguals were first exposed to English in a classroom context. In order to disentangle the potential influence of the FL context of learning from the onset age of FL acquisition, future research may investigate the MFLE in early and late bilinguals who learned their FL in the same immersive context (e.g., immigrants who arrived in the FL country at different ages; cf. Caldwell-Harris et al., 2012).

In addition to moral decision tendencies (for *Surgeon*), bilinguals' language background also modulated norm judgments (for *Surgeon* and *Bike week*) and emotional distress ratings (for *Factory* and *Bike week*). Bilingualism-related effects were not ubiquitous across scenarios, and were not always compatible with previous theorizing. For instance, in the face of a null effect of bilingual language experience on perceived emotional distress, the analysis of the *Surgeon* dilemma revealed a significant interaction between FL AoA and FL proficiency on norm judgment. In particular, deontological concerns decreased with increasing proficiency in early bilinguals, but increased with increasing proficiency in late bilinguals. The *increased deliberation* hypothesis states that lower proficiency in the FL leads to more deliberative cost-benefit appraisals that motivate harm acceptance to maximize outcomes. Put differently, increasing FL proficiency is predicted to amplify deontological inclinations. Whereas the permissibility ratings of late bilinguals are consistent with this prediction, those of early bilinguals showed an opposite pattern, which is in line with what we observed for the decision tendencies in the same dilemma.

The need to reexamine the scope of the previous conceptualizations of the FLE, as well as their implications on bilinguals' moral judgment, is particularly evident when looking at the permissibility ratings for *Bike week*. Here, a significant interaction between FL AoA and Language Dominance indicates that a moral violation becomes more acceptable with increasing exposure to the FL (or, more precisely, with decreasing asymmetry in dominance between a bilingual's two languages). This finding, similar in early and late bilinguals, is difficult to frame in any previous account of the MFLE.

Although we failed to corroborate the claim that the FLE works via attenuation of emotions (see Chan et al., 2016; cf. Gao et al., 2015; Hadjichristidis et al., 2015; Muda et al., 2020), we found that differences in the language profile of bilingual speakers affected their emotional reactions to moral conflict. In particular, a higher distress was predicted by increasing proficiency in the *Factory* dilemma, consistent with the hypothesis that high proficiency promotes emotional grounding (Costa et al., 2014). Moreover, a significant interaction was found between FL AoA and language

dominance in the *Factory* and *Bike week* dilemmas, indicating, in both cases, that increasing FL use diminishes emotional distance to the FL only in late bilinguals.

Overall, we provided evidence that differences in bilinguals' language background – conceptualized as the interplay of both static (i.e., FL AoA) and dynamic factors (i.e., FL proficiency and language dominance) – can influence moral decision-making in FL settings. At the same time, the characteristics of bilingual speakers, as well as their emotional reactions to moral conflict, are not necessarily responsible for the 'classic' MFLE, which may then depend on other factors when detected.

There are a number of limitations to the present study. One limitation is that we did not collect measures of decision-making style and personality traits from our participants, which may have moderated the observed effects or partially explained a lack thereof. Similarly, we did not collect measures of either dominance or proficiency in the NL condition. Another potential limitation is that we used sacrificial scenarios as experimental stimuli. Sacrificial scenarios have been criticized for low likelihood of occurrence (e.g., Bauman et al., 2014). Future research might therefore incorporate a wider variety of dilemmas that are also more representative of the moral situations people can face on an ordinary basis.

5. Conclusions

Although the nature of the FLE remains elusive, the contribution of this article to the extant literature is tangible in many respects. First, by using dilemmas other than the classic *Trolley/Bystander/Footbridge* scenarios, we provided evidence for a limited (i.e., stimulus-specific) effect of language context on participants' moral decisions. Second, asking for both willingness to act and action permissibility allowed us to support the conceptual division between moral decisions and norm judgments, suggesting that the framing of moral questions might be critical to elicit or not a MFLE. Third, by adopting a conceptualization of bilingualism that frames this construct as a composite and continuous measure, we were able to observe more nuanced and insightful effects of bilingual language experience on moral decision-making. To the extent that such effects were not consistent with previous theorizing, notably in the case of early bilinguals, we discussed the need to reconceptualize the FLE and its implications on moral decision-making.

Supplementary Materials. To view supplementary material for this article, please visit <http://doi.org/10.1017/langcog.2022.14>.

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, 569–576.
- Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108, 381–417.
- Baum, S., & Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics*, 35, 857–894.
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8, 536–554.
- Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, 85, 103855.

- Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, 18, 803–817.
- Brouwer, S. (2019). The auditory foreign-language effect of moral decision making in highly proficient bilinguals. *Journal of Multilingual and Multicultural Development*, 40, 865–878.
- Caldwell-Harris, C. L. (2015). Emotionality differences between a native and foreign language: Implications for everyday life. *Current Directions in Psychological Science*, 24, 214–219.
- Caldwell-Harris, C. L., Staroselsky, M., Smashnaya, S., & Vasilyeva, N. (2012). Emotional resonances of bilinguals' two languages vary with age of arrival: The Russian–English bilingual experience in the US. In P. Wilson (Ed.), *Dynamism in emotion concepts* (pp. 373–395). Peter Lang.
- Cao, F., Zhang, J., Song, L., Wang, S., Miao, D., & Peng, J. (2017). Framing effect in the trolley problem and footbridge dilemma: Number of saved lives matters. *Psychological Reports*, 120, 88–101.
- Carmona-Perera, M., Marti, C., Pérez-García, M., & Verdejo-García, A. (2013). Valence of emotions and moral decision-making: Increased pleasantness to pleasant images and decreased unpleasantness to unpleasant images are associated with utilitarian choices in healthy adults. *Frontiers in Human Neuroscience*, 7, 626.
- Čavar, F., & Tytus, A. E. (2018). Moral judgement and foreign language effect: When the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, 39, 17–28.
- Cecchetto, C., Rumiati, R. I., & Parma, V. (2017). Promoting cross-culture research on moral decision-making with standardized, culturally-equivalent dilemmas: The 4CONFiDe set. *Journal of Health and Social Sciences*, 22, 173–194.
- Chan, Y. L., Gu, X., Ng, J. C. K., & Tse, C. S. (2016). Effects of dilemma type, language, and emotion arousal on utilitarian vs deontological choice to moral dilemmas in Chinese-English bilinguals. *Asian Journal of Social Psychology*, 19, 55–65.
- Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology*, 5, 607.
- Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, 36, 1249–1264.
- Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign language effect. *Philosophical Psychology*, 29, 23–40.
- Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, 10, 216–235.
- Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, 179, 241–265.
- Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., & Keysar, B. (2017). Our moral choices are foreign to us. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1109–1128.
- Costa, A., Duñabeitia, J. A., & Keysar, B. (2019). Language context and decision-making: Challenges and advances. *Quarterly Journal of Experimental Psychology*, 72, 1–2.
- Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteigua, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS One*, 9, e94842.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12), 1082–1089.
- De Luca, V., Rothman, J., Bialystok, E., & Pliatsikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proceedings of the National Academy of Sciences*, 116, 7565–7574.
- 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*, 1–14. doi: 10.1017/S1366728921001012
- Fedeli, D., Del Maschio, N., Sulpizio, S., Rothman, J., & Abutalebi, J. (2021). The bilingual structural connectome: Dual-language experiential factors modulate distinct cerebral networks. *Brain and Language*, 220, 104978.
- Fricke, M., Zirnstein, M., Navarro-Torres, C., & Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, 22, 200–207.

- Garrigan, B., Adlam, A. L., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements. *Brain and Cognition*, 108, 88–97.
- Gawinkowska, M., Paradowski, M. B., & Bilewicz, M. (2013). Second language as an exemptor from sociocultural norms. Emotion-related language choice revisited. *PLoS One*, 8, e81225.
- Gao, S., Zika, O., Rogers, R. D., & Thierry, G. (2015). Second language feedback abolishes the “hot hand” effect during even-probability gambling. *Journal of Neuroscience*, 35(15), 5983–5989.
- Geipel, J., Hadjichristidis, C., & Surian, L. (2015a). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology*, 59, 8–17.
- Geipel, J., Hadjichristidis, C., & Surian, L. (2015b). The foreign language effect on moral judgment: The role of emotions and norms. *PLoS One*, 10, e0131529.
- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6, 517–523.
- Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364–371.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389–400.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105–2108.
- Gullifer, J. W., Chai, X. J., Whitford, V., Pivneva, I., Baum, S., Klein, D., & Titone, D. (2018). Bilingual experience and resting-state brain connectivity: Impacts of L2 age of acquisition and social diversity of language use on control networks. *Neuropsychologia*, 117, 123–134.
- Hadjichristidis, C., Geipel, J., & Savadori, L. (2015). The effect of foreign language in judgments of risk and benefit: The role of affect. *Journal of Experimental Psychology: Applied*, 21, 117–129.
- Haidt, J. (2007). The new synthesis on moral psychology. *Science*, 316, 998–1002.
- Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign-language effect on moral judgment. *Psychological Science*, 28, 1387–1397.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.
- Kant, I. (1959). *Foundations of the metaphysics of morals*. Macmillan.
- Kazanas, S. A., McLean, J. S., & Altarriba, J. (2019). Emotion and emotion concepts: Processing and use in monolingual and bilingual speakers. In J. W. Schwieter, & M. Paradis (Eds.), *The handbook of the neuroscience of multilingualism* (pp. 313–334). John Wiley & Sons.
- Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, 23, 661–668.
- Körner, A., Deutsch, R., & Gawronski, B. (2020). Using the CNI model to investigate individual differences in moral dilemma judgments. *Personality and Social Psychology Bulletin*, 46, 1392–1407.
- 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.
- Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, 72, 293–318.
- Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, 129(3), 361–368.
- Mill, J. S. (1998). Utilitarianism. In R. Crisp (Ed.), *Utilitarianism*. Oxford University Press.
- Mills, S., & Nicoladis, E. (2020). It’s easier to kill a baby to save oneself than a fat man to save other people: The effect of moral dilemma and age on Russian-English bilinguals’ moral reasoning. *Journal of Multilingual and Multicultural Development*, 1–10. doi: 10.1080/01434632.2020.1813145
- Miozzo, M., Navarrete, E., Ongis, M., Mello, E., Giroto, V., & Peressotti, F. (2020). Foreign language effect in decision-making: How foreign is it? *Cognition*, 199, 104245.
- Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, 6, 799–809.
- Montrul, S. (2016). Dominance and proficiency in early and late bilingualism. In C. Silva-Corvalan, & J. Treffers-Daller (Eds.), *Language dominance in bilinguals: Issues of measurement and operationalization* (pp. 15–35). Cambridge University Press.

- Muda, R., Niszczota, P., Bialek, M., & Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44, 321–326.
- Muda, R. P., Walker, A. C., Pieńkosz, D., Fugelsang, J. A., & Bialek, M. (2020). Foreign language does not affect gambling-related judgments. *Journal of Gambling Studies*, 36, 633–652.
- Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100, 530–542.
- O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision Making*, 5, 547.
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package 'mass'. *Cran r*, 538, 113–120.
- Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, 15(2), 165–184.
- R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Schrauf, R. W., & Rubin, D. C. (2000). Internal languages of retrieval: The bilingual encoding of memories for the personal past. *Memory & Cognition*, 28(4), 616–623.
- Slovan, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174–215.
- Stankovic, M., Biedermann, B., & Hamamura, T. (2022). Not all bilinguals are the same: A meta-analysis of the moral foreign language effect. *Brain and Language*, 227, 105082.
- Sulpizio, S., Del Maschio, N., Del Mauro, G., Fedeli, D., & Abutalebi, J. (2020). Bilingualism as a gradient measure modulates functional connectivity of language and control networks. *NeuroImage*, 205, 116306.
- Sulpizio, S., Del Maschio, N., Fedeli, D., & Abutalebi, J. (2020). Bilingual language processing: A meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 108, 834–853.
- Sulpizio, S., Toti, M., Del Maschio, N., Costa, A., Fedeli, D., Job, R., & Abutalebi, J. (2019). Are you really cursing? Neural processing of taboo words in native and foreign language. *Brain and Language*, 194, 84–92.
- Swann, W. B., Gómez, Á., Dovidio, J. F., Hart, S., & Jetten. (2010). Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science*, 21, 1176–1183.
- Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1), 35–39.
- Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, 4, 250.
- Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12, 579–590.
- Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4, 476–491.
- Wong, G., & Ng, B. C. (2018). Moral judgement in early bilinguals: Language dominance influences responses to moral dilemmas. *Frontiers in Psychology*, 9, 1070.

Cite this article: Del Maschio, N., Del Mauro, G., Bellini, C., Abutalebi, J. & Sulpizio, S. (2022). Foreign to whom? Constraining the moral foreign language effect on bilinguals' language experience *Language and Cognition* 14: 511–533. <https://doi.org/10.1017/langcog.2022.14>