Bilingualism: Language and Cognition

cambridge.org/bil

Research Article

Cite this article: Privitera, A. J., Li, S., Zhou, Y., & Wang, M. (2023). Modulatory role of foreign language experience on the Moral Foreign Language Effect. *Bilingualism: Language and Cognition*, **26**, 1038–1050. https://doi.org/10.1017/S1366728923000275

Received: 28 September 2022
Revised: 6 March 2023
Accepted: 14 March 2023
First published online: 22 May 2023

Keywords:
bilingualism; foreign language effect; moral decision-making; language proficiency; language dominance

Corresponding author:
Adam John Privitera, Email: aprivite@connect.hku.hk, (+65) 6908 1450 (mobile)

CAMBRIDGE
UNIVERSITY PRESS

Modulatory role of foreign language experience on the Moral Foreign Language Effect

Adam John Privitera[1,2], Shaohan Li[2], Yu Zhou[2] and Mengqi Wang[2]

[1]Centre for Research and Development in Learning, Nanyang Technological University, Singapore, Singapore and [2]College of Liberal Arts, Wenzhou-Kean University, Wenzhou, Zhejiang Province, China

Abstract

The Moral Foreign Language Effect (MFLE) is characterized by increased utilitarian decision-making when bilinguals respond to moral dilemmas in their foreign language. While previous research has given us a better understanding of this phenomenon, few studies have investigated how foreign language experience influences the MFLE. The present study investigated whether differences in foreign language proficiency, immersion, or dominance modulated the emergence of the MFLE. Mandarin–English bilingual young adults responded to a series of moral dilemmas in either their native or foreign language. Participants also provided ratings of permissibility and distress after reading each dilemma. We report a dilemma-specific MFLE that was modulated by differences in foreign language experience. Most significant was the observation that separable dimensions of foreign language experience interact when modulating the MFLE in a manner that is dilemma-specific. These findings emphasize the importance of considering differences in foreign language experience across multiple dimensions when investigating the MFLE.

Introduction

Modern globalization has created conditions under which bilinguals, people capable of using more than one language, may be expected to make decisions in their non-native language. Surprisingly, the decisions made by bilinguals are not stable across languages but differ based on the language in which the decision is made; a phenomenon dubbed the Foreign Language Effect (FLE). As originally described by Keysar, Hayakawa, and An (2012), the FLE is a phenomenon in which bilinguals demonstrate insensitivity to framing-effects when making decisions in their foreign language (FL) compared to their native language (NL). Over the last decade, the boundary conditions of this phenomenon have been investigated across a range of decision-making tasks in diverse samples of bilinguals around the world (Circi et al. 2021; Del Maschio et al. 2022b). Specifically, in the context of moral decision-making, the Moral Foreign Language Effect (MFLE) describes the tendency for bilinguals to make more utilitarian decisions when presented with moral dilemmas in their FL (Cipolletti et al., 2016; Corey et al. 2017; Costa et al. 2014; Stankovic et al., 2022). For example, in the classic *Footbridge* dilemma in which a person has the choice to save five people by pushing one person off a bridge and onto a track to block a runaway trolley (Thomson, 1985), bilinguals are more likely to decide to push the person when the dilemma is presented in their FL (Costa et al., 2014).

Two distinct systems are thought to underlie human decision-making: System I, a faster, more intuitive and emotional system; and System II, a slower, more deliberate system (Kahneman, 2003). In the context of moral decision-making (Greene & Haidt, 2002), engagement of System I is associated with deontological decisions in alignment with moral norms (e.g., not pushing the person off the bridge), while System II is associated with utilitarian decisions (e.g., sacrificing one person to save five). The MFLE is thought to result from decreased engagement of System I, and/or increased engagement of System II (Costa et al., 2019). Under both conditions, the presentation of a moral dilemma in an FL is expected to result in higher rates of utilitarian decisions, as well as higher ratings of permissibility – that is, whether or not the moral violation described in the dilemma is considered acceptable (Del Maschio et al., 2022b; Geipel et al., 2015a). However, evidence in support of increased engagement of System II (i.e., the increased deliberation hypothesis) is limited, with most accounts supporting decreased engagement of System I (e.g., Białek et al., 2019; Hayakawa et al. 2017).

Primary arguments for the decreased emotionality hypothesis state that the context in which an FL is generally learned, such as a classroom, "disembodies" the FL (Caldwell-Harris, 2014), decreasing the engagement of System I when processing moral dilemmas in an FL and consequently, increasing rates of utilitarian decisions (Costa et al., 2014). The

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

decreased emotionality hypothesis is further supported by the observation that bilinguals react less emotionally to offensive words presented in an FL compared to those in an NL (Dewaele, 2004), and report lower ratings of distress after reading a dilemma in an FL. In one study by Geipel et al. (2015a), participants assigned to the FL condition reported lower ratings of distress after reading both personal and impersonal versions of the *Trolley* dilemma. Further analysis revealed that distress was not a significant mediating factor on the MFLE, suggesting that reduced emotionality may manifest only in reported distress and not impact on decision-making. However, even evidence of decreased emotionality in self-reported distress ratings is mixed, with more recent investigations reporting null results in NL/FL group comparisons (Del Maschio et al., 2022b).

A number of past studies have investigated the robustness of the MFLE under various experimental conditions. Whether dilemmas are framed as personal or impersonal choices has been shown to influence the emergence of the MFLE. PERSONAL dilemmas require a participant to perform an action that will DIRECTLY kill another person (e.g., pushing a large man off of a bridge to prevent a trolley from killing a group of several workmen). Conversely, IMPERSONAL dilemmas require that an action be performed that INDIRECTLY kills another person (e.g., pulling a switch that will change the direction a trolley is traveling, resulting in the death of one person). Making a decision in an FL is thought to increase the psychological distance between the participant and the dilemma, increasing the likelihood of utilitarian responding to personal dilemmas (Costa et al., 2014), a prediction supported by a findings from a recent meta-analysis (Stankovic et al., 2022).

Expectedly, differences in language experience have also been shown to influence the MFLE. Costa et al. (2014) initially pointed to the influence of FL proficiency on the emergence of the MFLE, reporting that lower proficiency bilinguals were more likely to give utilitarian responses when dilemmas were presented in an FL. This finding, along with additional work supporting that the MFLE is less likely to emerge in bilinguals who are highly proficient in both languages (Brouwer, 2019) align with predictions generated from both THE INCREASED DELIBERATION and DECREASED EMOTIONALITY hypotheses (cf. Circi et al., 2021). As it follows, use of an FL by less proficient bilinguals is more cognitively demanding, and is therefore more likely to engage System II and result in HIGHER rates of utilitarian responding. In contrast, highly proficient bilinguals experience less cognitive load when using an FL and are less likely to engage System II, resulting in LOWER levels of utilitarian responding. Higher levels of FL proficiency may also increase how emotionally-grounded an FL is, increasing the possibility that System I is engaged, resulting in LOWER levels of utilitarian responding (Dewaele, 2004). However, it is likely the case that the FL learning context impacts on how strong the emotional connection is, possibly independent of proficiency (Costa et al., 2014).

The strength of conclusions that can be drawn from previous studies on the influence of linguistic variables on the MFLE is limited by methodological decisions. Most significant is the reliance on exclusively subjective assessments of FL proficiency. While subjective assessments are commonly used in studies on the effects of bilingualism (Anderson et al. 2018; Li et al. 2020; Marian et al., 2007), whether people can accurately assess their own abilities, including FL proficiency, is debated (MacIntyre et al., 1997; Zell & Krizan, 2014). Additionally, beyond self-reported assessment of FL proficiency, little else is generally known about the language experience of participants in these samples. This likely results from the lack of detailed assessments of language experience, a methodological issue across the wider literature on bilingualism (De Bruin, 2019). Although proficiency is the most widely-used metric for establishing bilingual language status (Surrain & Luk, 2017), the classification of bilingualism as a unidimensional construct lacks ecological validity and ignores the complex ways in which bilinguals differ across a number of separable dimensions (Gullifer et al. 2021).

Few studies have explored the influence of dimensions of language experience beyond proficiency on the MFLE. Limited evidence supports that more balanced NL/FL dominance (Wong & Ng, 2018) and higher linguistic similarity between the NL and FL (Dylman & Champoux-Larsson, 2020) are associated with LOWER levels of utilitarian responding when dilemmas are presented in an FL. Previous studies have also reported null results in the domain of immersion (Čavar & Tytus, 2018; Winskel & Bhatt, 2020), although it should be noted that samples in both studies were highly proficient. The cultural relevance of an FL has also been found to influence the emergence of the MFLE. In one study, Dylman and Champoux-Larsson (2020) reported that Swedish–English bilinguals did not demonstrate an MFLE on the *Footbridge* dilemma. In contrast, native Swedish speakers who were proficient in French, a less culturally influential language in Sweden, demonstrated an MFLE. This finding was attributed to the cultural influence of English in Sweden, due to the common use of English in music, films, and other popular emotionally-charged media. While investigations are limited, there is growing evidence that additional dimensions of language experience beyond proficiency may influence the MFLE.

Most recently, Del Maschio et al. (2022b) investigated the influence of FL age of acquisition (AoA; lowest age at which participants begin to listen to or learn to speak or write in an FL), objective FL proficiency (English Proficiency Test score), and FL dominance (aggregate score based on self-rated proficiency and hours of language use) on the MFLE in a sample of Italian–English bilinguals. In a departure from previous studies that generally ignore individual differences across participants, this investigation viewed bilingualism as a multidimensional, continuous construct. In their results, they reported dilemma-specific influences of separable dimensions of language experience on decision-making. While no MFLE was observed for the *Surgeon* dilemma (a personally-framed sacrificial dilemma adapted from the *Organ Transplant* dilemma; Cecchetto et al., 2017), early and late bilinguals differed in their decision-making based on their level of FL proficiency and FL dominance. Specifically, in early bilinguals, higher levels of FL proficiency were associated with higher rates of utilitarian responding only when FL dominance was low. In contrast, late bilinguals were less likely to give utilitarian responses when FL proficiency was high and FL dominance was low. Additionally, while early bilinguals were less likely to judge the moral violation described in the *Surgeon* dilemma as forbidden if their FL proficiency was higher, the opposite pattern was observed for late bilinguals. Finally, in the presence of a non-significant difference in ratings of distress between NL and FL groups, modulating influences of FL proficiency and FL dominance were identified on the *Factory* and *Bike Week* dilemmas (Cecchetto et al., 2017). Specifically, higher levels of FL proficiency were associated with higher ratings of distress on the *Factory* dilemma, and higher levels of FL dominance were associated with lower levels of distress on both the *Factory* and *Bike Week* dilemmas, but only in early bilinguals. Taken together,

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

these findings support that not only do separable dimensions of language experience impact on utilitarian responding and ratings of permissibility and distress, but that the effects of any one dimension are not identical across all bilinguals.

While the findings from Del Maschio et al. (2022b) highlight potentially complex interactions between separable dimensions of language experience on moral decision-making, this work must be considered in light of some limitations. Although both NL and FL groups were matched based on age, gender, and education, there is no evidence that groups were comparable in terms of language experience. For this reason, the results from between-groups analyses should be interpreted with some caution. Additionally, sacrificial dilemmas describing highly unlikely scenarios were used without context. While this is a common practice across previous studies, the use of instructions aimed at more appropriately contextualizing these unlikely scenarios may support more authentic responding (Christensen & Gomila, 2012). Dilemmas were also limited to only those classified as personal in the dimension of Personal Force (Christensen et al. 2014) preventing the investigation of whether the same pattern of results is observed on impersonal dilemmas. Finally, the use of AoA, operationalized as the lowest age a participant reported using an FL in any way (e.g., listening, reading, etc.) ignores additional ways in which bilinguals differ in their language experience such as years of use. When relying exclusively on the lowest age a participant began using an FL, a 20-year-old participant who began using an FL at age 3 and stopped at age 15 would be identical to a 45-year-old participant who began at age 3 and never stopped using an FL. Despite these limits, this study is significant in its contribution to our understanding of how separable dimensions of language experience modulate not only the MFLE, but ratings of permissibility (i.e., moral judgement) and perceived emotional distress, setting the stage for additional inquiry.

The present study aimed to further extend our limited understanding of how separable dimensions of FL experience modulate the MFLE in a sample of Mandarin–English bilinguals in China. Specifically, we investigated how differences in FL proficiency, FL immersion (an alternative to AoA which also captures differences in years of use), and FL dominance influenced utilitarian responding and ratings of permissibility and distress across a series of previously studied personal and impersonal moral dilemmas. In a departure from the majority of previous investigations, we matched experimental groups across a range of linguistic and non-linguistic background variables. We expected that participants would be more likely to make utilitarian decisions when presented with moral dilemmas in their FL, and that this effect would be most evident when dilemmas were framed from a personal perspective. Additionally, we expected that higher levels of FL proficiency would be associated with lower levels of utilitarian responding across dilemmas. Based on the observation that East Asian samples demonstrate lower levels of utilitarian responding when presented with moral dilemmas (Gold et al., 2014), and the view of both FL immersion and FL dominance as proxies for cultural experience (Čavar & Tytus, 2018; Winskel & Bhatt, 2020), we predicted higher rates of utilitarian responding associated with higher FL immersion and FL dominance. Finally, given the paucity of previous studies, our investigation of how differences in FL experience impact on ratings of permissibility and distress should be considered exploratory.

Our experiment was pre-registered: https://aspredicted.org/5WM_HRK. Data and materials for this experiment are available at https://doi.org/10.17605/OSF.IO/56YDZ.

Methods

Participants

A sample of 112 Mandarin–English bilinguals (81 females) was recruited from an English-immersive, Sino-American university located in Mainland China (M_{age} =20.90 years, $SD_{\text{age}} = 1.52$ years). This sample size exceeded the recommendation of $n = 101$ identified during an *a priori* power analysis conducted using G*Power version 3.1.9.6 (Faul et al. 2007) based on a previously reported effect size of $w = .28$ (Costa et al., 2014) for a significance threshold of $\alpha = .05$ and desired power = .80. All participants reported Mandarin as their native language (NL) and were enrolled in a fully English (FL) teaching environment. Written Informed consent was collected from all participants online prior to the start of the experiment. This study was approved by the Wenzhou-Kean University Institutional Review Board (#WKUIRB2022-006). Participants were randomly assigned to either the NL condition (i.e., Mandarin; $n = 62$, 74% female) or FL condition (i.e., English; $n = 50$, 70% female) in equal proportions based on reported sex at birth.

Assessing Language Experience

Because most previous studies on the MFLE limit the assessment of FL experience to only participants assigned to the FL condition, whether language condition groups are comparable across a range of relevant FL experience variables prior to analysis is rarely assessed. This is especially relevant when considering FL experience, particularly immersion, as a proxy for cultural experience, which is thought to influence moral decision-making (e.g., Costa et al., 2014; Gold et al., 2014; Winskel & Bhatt, 2020). To that end, all participants were asked to complete both subjective and objective assessments of language experience.

Subjective assessment

Participants completed the LHQ3 (Li et al., 2020) based on their experience with both Mandarin and English. The LHQ3 contains a series of self-report questions that assess separable dimensions of language experience for all languages that a person uses, resulting in separate aggregate scores ranging from 0 to 1. The use of LHQ3 aggregate scores allowed for bilingual experience to be operationalized as a multidimensional, continuous variable, aligning with recent calls to collect more detailed information about bilingual experiences (De Bruin, 2019), and an older call to avoid the assignment of categorical language status labels (Luk & Bialystok, 2013).

Proficiency in a given language is assessed via a single question for each of four components: listening, speaking, reading, and writing. Proficiency for each component is assessed using a 7-point Likert scale ranging from “very poor” to “excellent”. While there is some debate about the validity of self-reported measures of ability (e.g., Zell & Krizan, 2014), there is evidence to support that these measures correlate with objective ones (e.g., Gollan et al. 2011; Grant & Li, 2019). Immersion, the amount of time a person spends in a specific linguistic context, is calculated based on a participant’s age, years of language use, and the age at which they began listening, speaking, reading, and writing in a given language. Finally, dominance, the pattern of language use, is calculated based on proficiency and the reported number of hours listening, speaking, reading, and writing in a given language. Resulting aggregate scores for each dimension were weighted equally for each of the four components (i.e., 25% of the total aggregate score for listening, speaking, reading, and writing). While all three dimensions of language experience were assessed for English, Mandarin experience was limited to only self-reported proficiency. This decision was made based on our sample being drawn from a population of students who were born and raised in the Chinese Mainland, never living outside for an extended period of time. Additional questions were included on the LHQ3 in order to control for other variables that might underlie observed differences between groups including parental education level as a proxy for socioeconomic status (SES; Wermelinger et al., 2017) and estimated language switching frequency (How often are you in a situation when you switch between the languages Mandarin and English? 1 "never" to 7 "very often"). Finally, while subjective assessment of language experience was limited to only Mandarin and English, participants were asked to report which other languages they used in order to control for multilingualism.

Objective assessment

English (FL) proficiency was assessed objectively using an abbreviated version of an online, multiple-choice test developed by Transparent Language (test can be accessed for free online at https://www.transparent.com/). The version used in the present study included two sections taken from the source test: a 15-question section on English grammar, and a 10-question section on English vocabulary. Questions in both sections of the assessment were presented in the same order for all participants. Objective proficiency was operationalized as the total score across both sections of the assessment with each question weighted equally.

Materials

Participants were presented with three sacrificial moral dilemmas: the *Burning Building, Organ Transplant*, and *Trolley* dilemmas adapted from Christensen et al. (2014), a filler dilemma: *Train or Bus* (Geipel et al., 2015b), and two academic integrity dilemmas (data not reported in the present manuscript). Sacrificial moral dilemmas describe scenarios in which a person must decide whether to perform an action that will kill one person in order to save a larger number. Following the classification guidelines provided in the study by Christensen et al. (2014), all dilemmas were classified as avoidable in respect to Evitability, but varied across the dimensions of Personal Force (personal or impersonal), Benefit Recipient (self or other-beneficial), and Intentionality (instrumental or accidental). Classifications for each moral dilemma are presented in Table 1. In short: PERSONAL dilemmas are those that require that the decision-maker directly carry out the harmful action while IMPERSONAL involve the initiation of a process that leads to harm; SELF-BENEFICIAL dilemmas describe situations in which the decision-maker's life is at risk while OTHER-BENEFICIAL only involve the lives of others; finally, INSTRUMENTAL dilemmas are ones in which a harmful act is intentionally carried out for the greater good while ACCIDENTAL dilemmas describe scenarios in which the harm results as a side-effect.

Table 1. Classification of moral dilemmas based on the guidelines from Christensen et al. (2014). Personal Force is represented as PER (personal) and IMP (impersonal).

	Benefit Recipient	Evitability	Intentionality
Burning Building PER	Self	Avoidable	Instrumental
Burning Building IMP	Self	Avoidable	Accidental
Organ Transplant PER	Other	Avoidable	Instrumental
Organ Transplant IMP	Other	Avoidable	Instrumental
Trolley PER	Other	Avoidable	Instrumental
Trolley IMP	Other	Avoidable	Accidental

In total, participants responded to nine total dilemmas which included both personal and impersonal version of the three different moral dilemmas resulting in six total moral dilemmas. In both personal and impersonal dilemmas, choosing to commit the action (i.e., killing one person) is considered the UTILITARIAN DECISION while not committing the action is considered the DEONTOLOGICAL DECISION. Mandarin versions of the *Burning Building, Organ Transplant*, and *Trolley* dilemmas were taken from a study by Wong and Ng (2018). A non-moral filler dilemma (i.e., *Train or Bus*) was included in order to assess whether participants were able to understand the experimental task. This dilemma described a situation in which choosing to take a train from one city to another (yes response) would guarantee an on-time arrival to a meeting, while choosing the bus option (no response) may result in a late arrival. City names were changed from the original European locations to two major cities in China (i.e., Beijing and Jinan) in order to ensure that participant understanding was not impacted by the use of unfamiliar locations. A high percentage of "yes" responses was expected on this dilemma, with responses closer to 50% suggesting language comprehension issues. The Mandarin version of the non-moral filler dilemma was initially forward translated from English by a highly-proficient Mandarin–English bilingual researcher, checked by two independent bilingual researchers, and modified until consensus was reached. Mandarin and English versions of all dilemmas used in the present study can be found in **Supplementary Materials.**

Procedure

All tasks were administered using the Gorilla online experiment builder (Anwyl-Irvine et al. 2020). The decision to collect data through an Internet-based platform was made to avoid complications associated with future local outbreaks of COVID-19 that might otherwise prevent the collection of in-person data. After participants clicked on the experimental task link, they were automatically assigned a non-identifying alphanumeric ID number and were screened based on their access device and self-reported native language. Participants who accessed the experiment using a phone or tablet, or those reporting anything other than Mandarin as their native language were automatically rejected. Participants were then automatically grouped by reported sex at birth to allow for equal proportions of male and female participants to be assigned to the NL and FL conditions. Task order was identical across all participants.

Participants were first presented with the study consent form written in the language of the experimental group they were assigned to. Participants were asked to check a box and type their name if they agreed to participate in the study. Completed consent forms were downloaded from the online experiment builder and stored separately from all other data. Previous studies conflict regarding whether consent forms should be presented in the same language of the assigned language condition (e.g., Bereby-Meyer et al. 2020) or in participants' native language to

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

ensure comprehension (e.g., Dylman & Champoux-Larsson, 2020). Consent form comprehension issues in the FL condition were not expected due to the use of non-technical language and the linguistic background of our sample.

Next, participants completed the dilemma task in the language of their assigned experimental condition, including all directions and on-screen buttons. The translation of directions from English to Mandarin was accomplished in same manner as the non-moral filler dilemmas. Given that the situations described in each of the moral dilemmas were unlikely to ever be encountered in real life, and because personal and impersonal versions of the same dilemma were highly similar, participants were provided with directions based on guidance from Christensen and Gomila (2012) that helped to better contextualize the dilemmas and prepare participants: *"In the following you will read a series of short stories about difficult situations, similar to those that you may have seen in the news or in a book or movie. Although these stories might seem similar, they are different in important ways" / "在下文中，你将读到一系列关于比较棘手的问题的短篇故事，可能与你在新闻或书本或电影中看到的情况类似。尽管这些故事可能看起来很相似，但它们在重要方面是不同的。".* Participants were also asked to maximize the size of their Internet browser in order to prevent distractions during the experiment. To reduce the potential influence of dilemma order, all nine dilemmas were presented randomly, one at a time, in centered, 20-point black font on a white background. Unlimited time was provided in order to read each dilemma and to give either a "yes" or "no" response to whether the participant would perform the described action.

For each dilemma presented, after selecting either the "yes" or "no" response, a new screen appeared, and participants were asked to rate the permissibility of the action described in the dilemma on a 7-point Likert scale: *How do you rate this action? (rate from 1 - 7) / 你会如何给这种行为打分？（从1-7进行评分).* Verbal anchors were provided on both the left side closest to 1: "*This action should never be done" / "这种行为永远不应该做"*, and the right side closest to 7: *"It is necessary for this action to be done" / "这样做是有必要的".* On the next screen, participants were asked to rate their distress about the dilemma: *"Thinking about the scenario I just read, I feel very troubled (rate from 1 - 7)" / "想到刚才我读的那个场景，我感觉非常困扰（从1-7进行评分）"*, with verbal anchors provided on the left side: "Strongly Disagree" / *"非常不同意"*, and right side: *"Strongly Agree" / "非常同意".* On the final screen, participants were asked to rate their understanding of the dilemma: *"Did you understand the story you just read? (rate from 1 - 7)" / "你理解你刚才读的那个故事吗？（从1-7进行评分）"*, with verbal anchors on the left: *"Not at All" / "一点也不"* and right side: *"Very Well" / "非常好".* Participants had an unlimited amount of time in order to answer questions for each dilemma. Question order was identical for each dilemma across all participants (i.e., permissibility, distress, and comprehension).

After the completion of the dilemma task, participants completed a bilingual version of the LHQ3. The collection of language experience data occurred after the completion of the experimental dilemma block in order to prevent the identification of language as a variable of interest from influencing performance on the dilemmas. Because the LHQ3 was administered after the completion of the experimental dilemmas, a bilingual version was considered acceptable in the interest of ensuring that nuanced questions about language experience were clearly understood. Finally, after completing a two-color Simon task (data not reported in the present manuscript) participants completed the objective FL proficiency assessment with no time limit, and were then debriefed. In total, the experiment took around 20 minutes to complete for both experimental groups.

Data analysis

All statistical analyses were performed in R (R Core Team, 2021). Given the nature of our experimental design, and our specific research questions, both between-groups and within-group analyses were used. Prior to analysis, data from individual dilemmas in which a participant reported a comprehension score of 3 or lower were excluded.

Between-groups analysis

In order to test the impact of language condition and dilemma type on decision-making, separate binomial logistic regression models with response (yes/no) as the dependent variable and language condition (NL vs. FL) and dilemma type (personal vs. impersonal) as predictors were run for each dilemma using the glm function from the stats package (Version 4.0.5). Differences in rated permissibility and distress for each dilemma were tested using separate ordinal logistic regressions models with Likert-scale ratings for permissibility and distress as the dependent variable and language condition (NL vs. FL) and dilemma type (personal vs. impersonal) as predictors using the polr function from the MASS package (Version 7.3-54; Venables & Ripley, 2013). For all models, interaction terms were initially included for language condition and dilemma type and were removed if they did not improve model fit.

Within-group analysis

To assess the influence of separable dimension of FL experience on decision-making in our FL condition, separate binomial logistic regression models with response (yes/no) as the dependent variable and FL proficiency, FL immersion, FL dominance, and all two-way interactions between as predictors were run for each dilemma using the glm function from the stats package (Version 4.0.5). Additionally, to test whether differences in FL experience impacted on ratings of permissibility and distress, separate ordinal logistic regression models with Likert-scale ratings for permissibility and distress as the dependent variable and FL proficiency, immersion, and dominance as predictors were run for each dilemma using the polr function from the MASS package. All predictors were standardized before analysis. For each model, forward and backward stepwise regression were performed in order to find the best fitting model using the stepAIC function from the MASS package. Final model selection was confirmed based on manual selection of the model with the lowest Akaike information criterion (AIC; Stoica & Selen, 2004). Generally, a reduction of ≥ 2 AIC units is considered a significant improvement in goodness of fit when comparing models (Anderson & Burnham, 2004).

Results

Application of the comprehension exclusion criteria resulted in the removal of data from eight participants for the *Burning Building* dilemma (six personal version), seven participants for the *Organ Transplant* dilemma (six personal version), six participants for the *Trolley* dilemma (three personal version), and five participants for the *Train or Bus* filler dilemma. Additionally,

Table 2. Demographic and language experience background of participants. NL: native language; FL: foreign language; SUB: subjective; OBJ: objective. Reported p-values were generated from chi-square test of association (percentage female) and Welch's t-test (all other variables).

	NL Condition ($n = 58$)	FL Condition ($n = 48$)	p value
Percentage female	76%	73%	0.729
Age (years)	20.81 (1.26)	20.79 (1.27)	0.940
Years of education	15.05 (1.13)	15.29 (1.07)	0.113
SES (1-6 points)	3.62 (0.92)	3.66 (0.73)	0.825
Age of acquisition	7.64 (3.29)	7.35 (2.11)	0.592
FL experience (years)	13.17 (3.41)	13.44 (2.38)	0.640
NL proficiency SUB (0-1 point)	0.77 (0.13)	0.80 (0.14)	0.271
FL proficiency SUB (0-1 point)	0.60 (0.11)	0.59 (0.14)	0.391
FL proficiency OBJ (0-1 point)	0.81 (0.10)	0.79 (0.09)	0.878
FL immersion (0-1 point)	0.64 (0.12)	0.64 (0.09)	0.938
FL dominance (0-1 point)	0.36 (0.08)	0.34 (0.09)	0.347
Language switching (1-7 points)	4.17 (1.42)	4.38 (1.79)	0.527

five participants were removed due to having recorded response times of less than 1000 ms for any single dilemma. This prevented the inclusion of participants who did not read a dilemma prior to making a response. Finally, one participant was removed due to reporting an age that was significantly above the rest of the sample. This resulted in the inclusion of data from 106 participants (79 females) in our analyses. The proportion of females in both language conditions was roughly equivalent after pre-analysis trimming (NL: 76%; FL: 73%) and did not differ significantly ($X^2(1, n = 106) = 0.120$, $p = .729$). Demographic characteristics of our sample are reported in Table 2. Distributions of FL experience variables were visual inspected, confirming that scores were sufficiently distributed across each dimension prior to analysis. For all models, reference levels for categorical variables were set as follows: response ("no" as reference level), language condition ("NL" as reference), and dilemma type ("personal" as reference).

Between-groups analysis: influence of language condition and dilemma type

Language condition groups did not differ significantly based on age, education, SES, subjective FL proficiency, objective FL proficiency, FL immersion, FL dominance, subjective NL proficiency, or years of FL experience ($ps \geq .113$). Therefore, language condition groups were considered equivalent prior to analysis. The percentage of yes responses (i.e., utilitarian responses), and average ratings for permissibility and distress are presented for both versions (i.e., personal or impersonal) of each moral dilemma, disaggregated by language condition in Table 3. For all between-groups analysis models, the inclusion of the language condition and dilemma type interaction term did not significantly improve model fit. For this reason, interaction terms were not included in any models. No significant main effect of language condition was observed for the *Train or Bus* filler dilemma on utilitarian responding ($X^2(1) = 0.135$, $p = .713$) or ratings of permissibility or distress ($ps \geq .319$)

Burning Building dilemma

A significant main effect of language condition was observed ($X^2(1) = 6.069$, $p = .015$, $\beta = -0.805$, $SE = .331$, $z = -2.434$), supporting a REVERSE MFLE on this dilemma. Specifically, participants were more likely to give a utilitarian response when the *Burning Building* dilemma was presented in their NL. Additionally, a significant main effect of dilemma type was observed ($X^2(1) = 5.232$, $p = .024$, $\beta = 0.751$, $SE = .333$, $z = 2.256$), with participants more likely to give a utilitarian response when presented with the impersonal version of this dilemma. For permissibility ratings, the overall model test was marginally significant ($X^2(2) = 4.718$, $p = .095$), with a significant effect of dilemma type ($X^2(1) = 4.358$, $p = .038$, $\beta = 0.519$, $SE = .249$, $z = 2.080$). Participants gave higher ratings of permissibility for the impersonal version of this dilemma. Finally, the overall model for ratings of distress was not significant ($X^2(2) = 1.583$, $p = .453$).

Organ Transplant dilemma

A significant main effect of language condition was observed ($X^2(1) = 5.838$, $p = .022$, $\beta = 1.253$, $SE = .549$, $z = 2.283$), with a higher likelihood of a utilitarian response when the dilemma was presented in the FL, supporting an MFLE. No significant main effect of dilemma type was observed for utilitarian responding ($X^2(1) = 1.923$, $p = .165$). Finally, no significant main effects of language condition or dilemma type were found for ratings of permissibility or distress ($ps \geq .176$).

Trolley dilemma

No significant main effect of language condition was observed for the *Trolley* dilemma ($X^2(1) = 0.432$, $p = .511$). A significant main effect of dilemma type was observed ($X^2(1) = 21.602$, $p < .001$, $\beta = 1.628$, $SE = .379$, $z = 4.300$), with participants more likely to give a utilitarian response when presented with the impersonal version of the dilemma. Additionally, when modeling reported permissibility, the overall model was significant ($X^2(2) = 11.590$, $p = .003$), with a significant main effect of dilemma type ($X^2(1) = 11.422$, $p < .001$, $\beta = 0.850$, $SE = .254$, $z = 3.347$). Participants gave higher ratings of permissibility when presented with the impersonal version of the dilemma. Finally, when modeling reported distress, a marginally significant overall model test was found ($X^2(2) = 5.639$, $p = .060$), with a significant effect of dilemma type ($X^2(1) = 2.289$, $p = .022$, $\beta = 0.569$, $SE = .248$, $z = 2.289$). Participants reported higher levels of distress after reading the impersonal version of the dilemma.

Within-Group Analysis: Influence of Separable Dimensions of FL Experience

Due to high variance inflation factor between subjective FL proficiency and FL dominance, objective FL proficiency scores were substituted for subjective scores in each model. This decision addressed issues with multicollinearity while also supporting the inclusion of three separable dimensions of FL experience. After substitution, variance inflation factor levels were at acceptable levels for all models (< 5; Craney & Surles, 2002). The use of objective FL proficiency scores resulted in the removal of an additional eight participants from the FL condition who did not complete the assessment ($n = 46$; 33 females). Models for the *Train or Bus* filler dilemma were not significant for utilitarian responding ($X^2(1) = 1.528$, $p = .216$) or ratings of permissibility or distress ($ps \geq .752$).

Table 3. Percentage of yes responses (i.e., utilitarian) and average ratings of permissibility and distress for each dilemma. Standard deviations are reported in parentheses.

	% Yes (NL)	% Yes (FL)	Permiss (NL)	Permiss (FL)	Distress (NL)	Distress (FL)
Burning Building PER	76%	57%	4.00 (1.78)	4.09 (1.74)	5.11 (1.89)	5.35 (1.84)
Burning Building IMP	86%	75%	4.39 (1.79)	4.65 (1.67)	4.82 (2.12)	5.17 (1.77)
Organ Transplant PER	2%	11%	2.36 (1.89)	2.73 (2.27)	3.51 (2.33)	3.51 (2.19)
Organ Transplant IMP	7%	17%	2.86 (2.10)	2.98 (2.06)	3.51 (2.19)	4.13 (2.09)
Trolley PER	7%	15%	2.91 (1.79)	3.02 (1.82)	4.23 (2.24)	4.24 (2.07)
Trolley IMP	38%	38%	3.80 (1.93)	3.83 (1.91)	4.69 (2.33)	5.17 (1.74)
Train or Bus	97%	98%	6.11 (1.13)	5.71 (1.58)	2.75 (1.68)	2.89 (2.09)

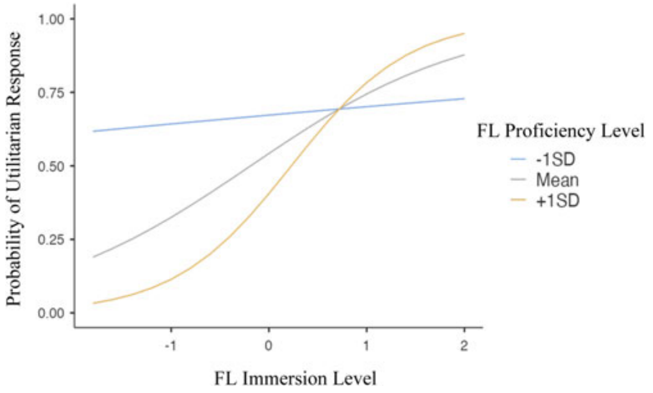

Figure 1. Influence of language experience on utilitarian responding in the personal version of the *Burning Building* dilemma. Marginally significant interaction between FL immersion and FL proficiency. Data are from participants in the foreign language condition (n = 46).

Burning Building dilemma: personal version.

For the personal version of the *Burning Building* dilemma, a significant overall model test was identified for a model containing FL proficiency, FL immersion, and their interaction ($X^2(3) = 12.810$, $p = .005$). A significant effect of FL immersion was observed with higher levels of FL immersion associated with higher rates of utilitarian responding ($X^2(1) = 5.619$, $p = .026$, $\beta = 0.850$, $SE = 0.383$, $z = 2.220$). This main effect should be considered in light of a marginally significant interaction between FL immersion and FL proficiency ($X^2(1) = 3.704$, $p = .075$) where the influence of FL immersion was initially most pronounced at the lowest levels of FL proficiency until higher levels of FL immersion were reached in more proficient bilinguals (**Figure 1**). When modeling ratings of permissibility, a marginally significant overall model test was observed for the personal version of the dilemma for a model containing main effects for FL proficiency, FL immersion, and FL dominance along with the interaction between FL immersion and FL dominance ($X^2(4) = 8.898$, $p = .064$). A marginally significant interaction between FL dominance and FL immersion was identified ($X^2(1) = 3.745$, $p = .058$, $\beta = -0.748$, $SE = 0.394$, $z = -1.898$). To better understand the nature of this interaction, FL immersion was median split and recoded into "Low Immersion" and "High Immersion" levels. As shown in **Figure 2**, while higher reported levels of FL dominance were associated with higher ratings of permissibility on the personal version of the dilemma, this pattern was only observed for those with low FL immersion. Higher levels of FL dominance associated with slightly higher ratings of permissibility were observed in those with high FL immersion.

Burning Building dilemma: impersonal version

When modeling reported permissibility, a significant overall model test was observed for the impersonal version of this dilemma for a model containing main effects for FL proficiency and FL dominance along with their interaction ($X^2(3) = 9.381$, $p = .025$). A significant interaction between FL proficiency and FL dominance was identified ($X^2(1) = 7.704$, $p = .009$, $\beta = 1.012$, $SE = 0.386$, $z = 2.751$). To better understand the nature of this interaction, FL proficiency was median split and recoded into "Low Proficiency" and "High Proficiency" levels. As shown in **Figure 3**, while higher reported levels of FL dominance were associated with lower ratings of permissibility, this pattern was only observed for those with low FL proficiency. The opposite pattern, higher levels of FL dominance associated with higher ratings of permissibility, was observed in those with high FL proficiency.

Organ Transplant dilemma: personal version

A significant overall model test was observed for the personal version of the *Organ Transplant* dilemma for a model containing only FL proficiency ($X^2(2) = 6.870$, $p = .009$). Higher levels of FL proficiency were associated with decreased utilitarian responding ($X^2(1) = 6.870$, $p = .025$, $\beta = -1.378$, $SE = 0.614$, $z = -2.242$). When modeling reported permissibility, a significant overall model test was observed for the personal version of this dilemma for a model containing main effects for FL proficiency and FL immersion along with their interaction ($X^2(3) = 8.882$, $p = .031$). A significant interaction between FL proficiency and FL immersion was identified ($X^2(1) = 4.754$, $p = .040$, $\beta = 0.859$, $SE = 0.419$, $z = 2.050$). To better understand the nature of this interaction, FL immersion was median split and recoded into "Low Immersion" and "High Immersion" levels. As shown in **Figure 4**, while higher reported levels of FL proficiency were associated with lower ratings of permissibility on the personal version of the dilemma, this pattern was only observed for those with low FL immersion. The opposite pattern, higher levels of FL proficiency associated with slightly higher ratings of permissibility, was observed in those with high FL immersion.

Organ Transplant dilemma: impersonal version

When modeling permissibility ratings, a marginally significant overall model test was observed for the impersonal version of the dilemma for a model containing a main effect of FL

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

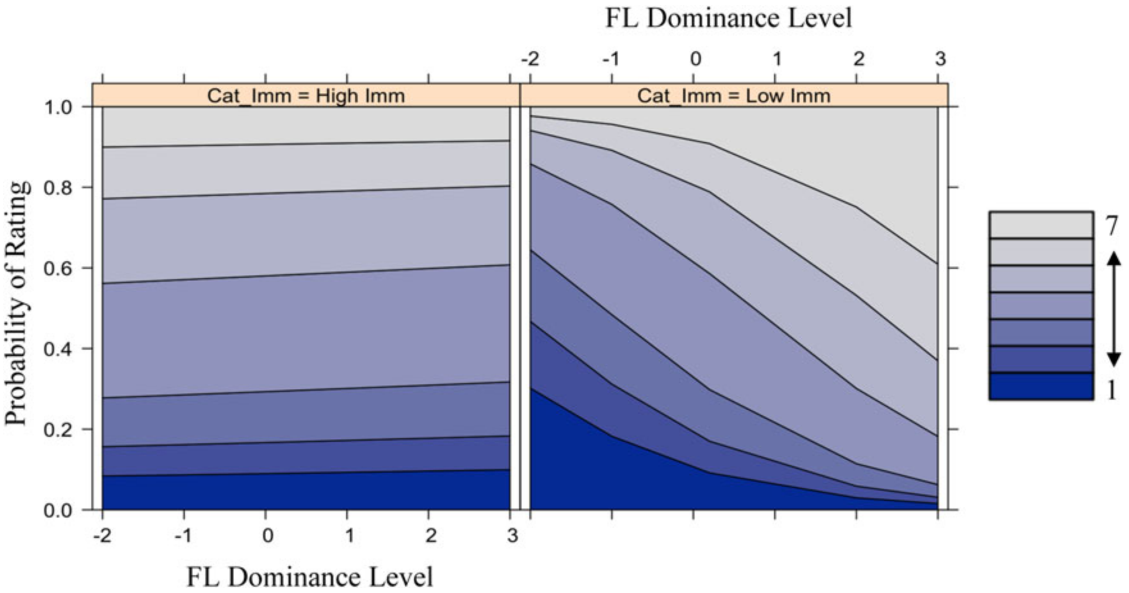

Figure 2. Influence of language experience on permissibility ratings in the personal version of the *Burning Building* dilemma. Marginally significant interaction between FL immersion and FL dominance. Data plotted separately for High Immersion (Left) and Low Immersion (Right) groups. Lighter colors represent higher ratings of permissibility. Data are from participants in the foreign language condition (*n* = 46). Language experience levels are standardized.

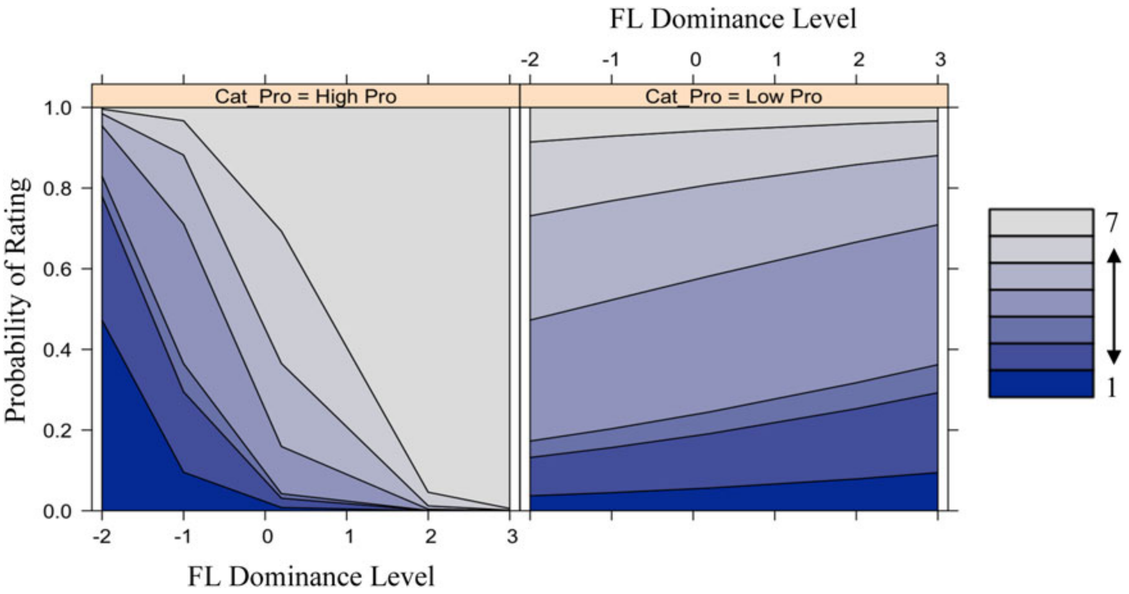

Figure 3. Influence of language experience on permissibility ratings in the impersonal version of the *Burning Building* dilemma. Significant interaction between FL proficiency and FL dominance. Data plotted separately for High Proficiency (Left) and Low Proficiency (Right) groups. Lighter colors represent higher ratings of permissibility. Data are from participants in the foreign language condition (*n* = 46). Language experience levels are standardized.

proficiency ($X^2(1) = 2.868$, $p = .090$). A marginally significant main effect of FL proficiency was observed with higher levels associated with decreased ratings of permissibility ($X^2(1) = 2.868$, $p = .098$, $\beta = -0.494$, $SE = 0.298$, $z = -1.657$).

Trolley dilemma: personal version

When modeling reported distress, a significant overall model test was observed for the personal version of this dilemma for a model containing main effects for FL proficiency, FL immersion, and FL dominance along with the interaction between FL immersion and FL dominance ($X^2(4) = 11.175$, $p = .025$). A significant interaction between FL immersion and FL dominance was identified ($X^2(1) = 4.191$, $p = .045$, $\beta = -0.768$, $SE = 0.382$, $z = -2.008$). To better understand the nature of this interaction, FL dominance was median split and recoded into "Low Dominance" and "High Dominance" levels. As shown in **Figure 5**, while higher reported levels of FL immersion were associated with higher ratings of distress on the personal version of the dilemma, this pattern was only observed for those with low FL dominance. The opposite pattern, higher levels of FL immersion associated with lower

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

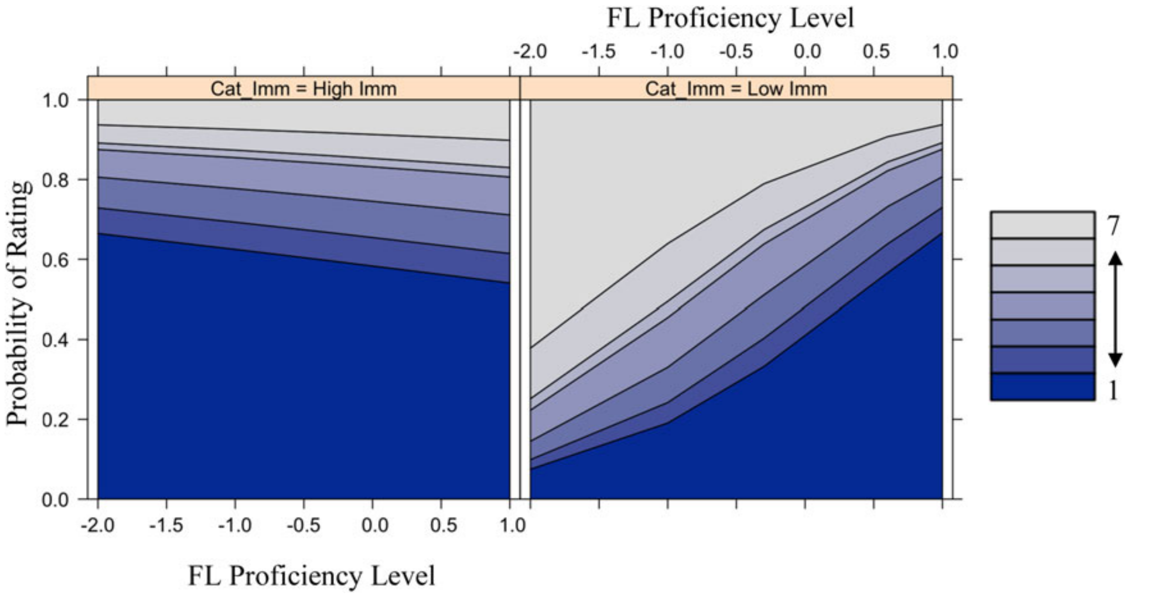

Figure 4. Influence of language experience on permissibility ratings in the personal version of the *Organ Transplant* dilemma. Significant interaction between FL proficiency and FL immersion. Data plotted separately for High Immersion (Left) and Low Immersion (Right) groups. Lighter colors represent higher ratings of permissibility. Data are from participants in the foreign language condition ($n = 46$). Language experience levels are standardized.

ratings of distress, was observed in those with high FL dominance. Additionally, a marginally significant main effect of FL proficiency was observed, with higher levels associated with lower ratings of distress ($X^2(1) = 3.465$, $p = .067$, $\beta = -0.573$, $SE = 0.313$, $z = -1.833$).

Trolley dilemma: impersonal version

No overall model test for any variable investigated was significant for this version of the dilemma ($ps \geq .266$).

Discussion

We report a dilemma-specific MFLE and partial evidence for modulatory influences of dilemma framing and FL experience on that effect. As summarized in Table 3, an MFLE was only observed on the *Organ Transplant* dilemma, while a reverse MFLE was observed on the *Burning Building* dilemma. Additionally, participants were less likely to make utilitarian decisions on the personal versions of both the *Burning Building* and *Trolley* dilemmas regardless of the language they were presented in, a finding that was also observed in ratings of permissibility. We also report limited evidence in support of the influence of FL experience on decision-making in the FL condition with significant effects of FL proficiency and FL immersion observed on the personal versions of the *Organ Transplant* and *Burning Building* dilemmas, respectively. Finally, separable dimensions of FL experience, often interacting with each other, impacted on ratings of permissibility and, in the personal version of the *Trolley* dilemma only, ratings of distress. This work adds to a growing literature investigating graded effects of separable dimensions of language experience on previously reported psycholinguistic phenomena (Privitera et al., 2022a, 2022b).

We did not observe an MFLE consistently across all dilemmas, replicating the null and mixed findings of some previous studies (Čavar & Tytus, 2018; Costa et al., 2014; Del Maschio et al., 2022b; Wong & Ng, 2018). In fact, an MFLE was only observed on the *Organ Transplant* dilemma. The *Organ Transplant* and related *Surgeon* and *Transplant* dilemmas are not widely used across previous studies (Del Maschio et al. 2022a). To the best of our knowledge, only one study has reported a significant MFLE on this dilemma (Shin & Kim, 2017) with most reporting null results (Čavar & Tytus, 2018; Del Maschio et al., 2022b; Wong & Ng, 2018) although null results are contentious (Białek & Fugelsang, 2019). In the *Organ Transplant* dilemma, the participant is taking the perspective of a doctor who must make a choice that violates the norms associated with that role (i.e., causing harm to a patient). In contrast, no other dilemmas utilized in the present study placed the participant in a role with explicit norms against causing harm. For this reason, the moral violation in the *Organ Transplant* dilemma may have been more salient than those described in other dilemmas. The observation of an MFLE on this dilemma aligns with reduced emotionality, increased deliberation, and reduced access to social norm models (Geipel et al., 2015a; Keysar et al., 2012). Interestingly, we did not report any difference in ratings of distress between NL and FL conditions on this dilemma. This may reflect the insensitivity of our measure to differences in emotional experience generated from reading moral dilemmas, a finding that has been previously reported in a similar dilemma (Del Maschio et al., 2022b).

Surprisingly, we observed a reverse MFLE on the *Burning Building* dilemma and no MFLE on the widely-studied *Trolley* dilemma. A reverse MFLE on the *Burning Building* conflicts with previously reported null results (Wong & Ng, 2018). This dilemma is unique compared to the other dilemmas utilized in the present study as it is the only self-beneficial dilemma, describing a scenario in which the participant would die if they did not commit the moral violation. To the best of our knowledge, no other study has reported a reverse MFLE on the *Burning Building* dilemma. One possible explanation for this observed result could relate to an increased sensitivity to negative emotions when self-beneficial dilemmas are presented in the NL. Self-beneficial dilemmas may also be considered less emotional

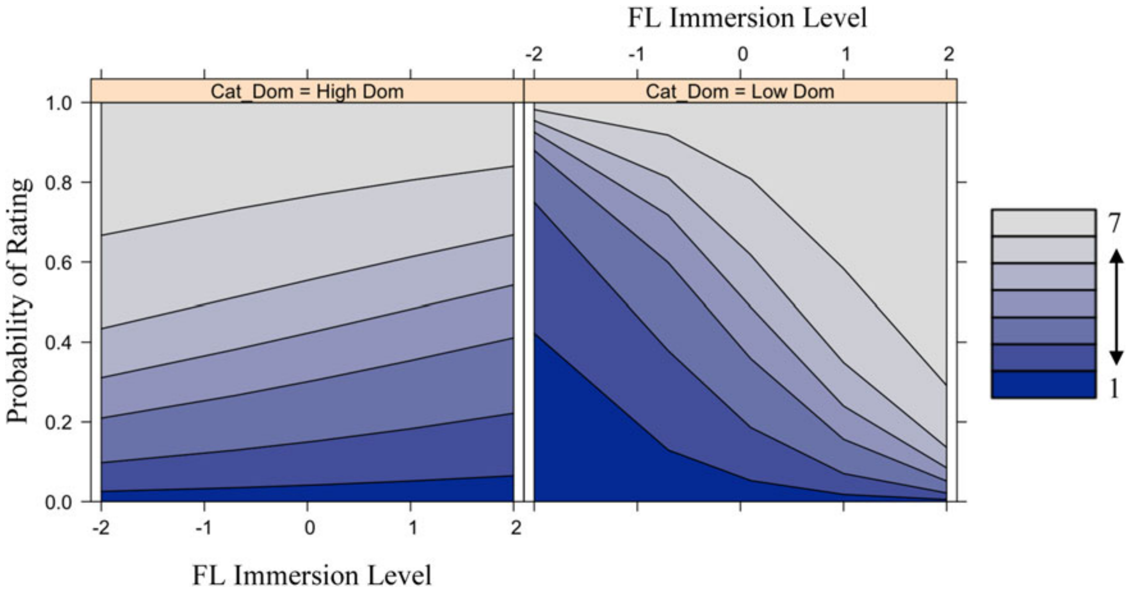

Figure 5. Influence of language experience on distress ratings in the personal version of the *Trolley* dilemma. Significant interaction between FL immersion and FL dominance. Data plotted separately for High Dominance (Left) and Low Dominance (Right) groups. Lighter colors represent higher ratings of permissibility. Data are from participants in the foreign language condition ($n = 46$). Language experience levels are standardized.

when presented in an FL (Costa et al., 2014), further reducing the likelihood that participants will respond in the interest of self-preservation. Lack of an MFLE on the *Trolley* dilemma is more difficult to explain. We speculate that this null result may relate to participant familiarity as both versions of this dilemma are widely depicted in popular culture and discussed in a number of different high school and college-level courses. Considering that all participants completed high school in Mainland China and were presently enrolled in an English-immersive university, we expect that this dilemma was likely familiar in both Mandarin and English. Unexpectedly, participants gave higher distress ratings on the impersonal version of the *Trolley* dilemma. This result is difficult to interpret, but may stem from the emotionally-dampening influence of repeated exposure to the personal version of this dilemma, an experience that has been shown to desensitize personal reactions to shocking stimuli (Campbell et al. 2014). Future research should consider assessing dilemma familiarity in order to explore this question directly.

Despite our mixed findings regarding the MFLE on the *Burning Building* and *Trolley* dilemmas, we did observe the previously reported increase in utilitarian responding when participants were presented with the impersonal version of each scenario across both language conditions, supporting that differences between versions were salient (Costa et al., 2014; Wong & Ng, 2018). Evidence in support of a preference for indirect harm is not unique to studies utilizing written moral dilemmas (e.g., Milgram, 1974) and is thought to result from the influence of an evolved aggression-inhibitory mechanism that generates a sense of moral wrongness in the presence of cues related to causing physical harm to another person (Royzman & Baron, 2002). Replication of this finding across diverse samples of participants that differ by age, gender, and nationality provides further support for this evolutionary argument (Hauser et al. 2007). This preference was also observed in ratings of permissibility for both dilemmas, supporting that the influence of this mechanism may extend beyond personal decisions into the domain of moral judgements. Notably, a preference for indirect harm was not observed on the *Organ Transplant* dilemma, conflicting with previous reports (e.g., Wong & Ng, 2018). The moral incongruence of medical professionals causing harm either directly or indirectly is evidenced in bioethical debates about the participation of physicians in prisoner executions (Black & Fairbrother, 2008). Clear violation of the "first do no harm" principle of the widely-known Hippocratic Oath may be responsible for the observed low levels of utilitarian responding on both versions of this dilemma, preventing the identification of a significant effect of dilemma type. Alternatively, while our sample size was comparable to the majority of past studies on both the FLE in general and the MFLE specifically (Del Maschio et al., 2022a), a significant effect of dilemma type may have emerged with a larger sample. This is further supported by our observed trend toward higher rates of utilitarian responding on the impersonal version of the *Organ Transplant* dilemma ($p = .176$).

Modulatory influences of separable dimensions of language experience on utilitarian decision-making in the FL condition were only present on the personal versions of the *Burning Building* and *Organ Transplant* dilemmas. While the predicted decrease in utilitarian responding associated with higher levels of FL proficiency was observed on the *Organ Transplant* dilemma, a marginally significant interaction between FL proficiency and FL immersion was observed on the *Burning Building* dilemma. Higher FL proficiency has been shown to lead to increased levels of emotionality (e.g., Caldwell-Harris, 2015) and decreased levels of cognitive load associated with FL use (e.g., Hayakawa et al., 2017). Under either of these conditions, a higher likelihood of intuitive System I activation or lower activation of deliberate System II would result in lower levels of utilitarian responding, leading to the absence of an MFLE in highly proficient bilinguals (Brouwer, 2019). Higher FL proficiency was also associated with lower levels of utilitarian responding on the *Burning Building* dilemma, however, this influence was not observed at the highest levels of FL immersion. We interpret these findings as consistent

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

with dual-system theories (Kahneman, 2011) but highlight a modulating effect of FL immersion.

Few studies have examined the influence of FL immersion on the MFLE with extant studies reporting null results (Čavar & Tytus, 2018; Winskel & Bhatt, 2020). While more immersed bilinguals tend to be more proficient (Kinsella & Singleton, 2014), this was not the case with our sample *r*(98) = .069, *p* = .497, and may reflect a unique linguistic characteristic of students enrolled in an FL-immersive university located in an NL-immersive country. Higher levels of FL immersion may also reflect higher levels of acculturation (Čavar & Tytus, 2018). Based on the previously-reported bias toward deontological choices in East Asian samples (Gold et al., 2014), higher levels of immersion in English would be expected to increase utilitarian responding due to heightened sensitivity to the social norms of English language culture (Geipel et al., 2015a). However, as reported, the effects of acculturation through FL immersion were diminished at higher levels of FL proficiency unless FL immersion was sufficiently high, suggesting that sensitivity to norms may increase cognitive load and increase engagement of System II. One unexplored alternative is that this result was not due to increased sensitivity to social norms but actually resulted from stronger identity with Western culture due to high levels of FL immersion. Whether FL immersive schooling in an NL immersive country significantly impacts cultural identity is an understudied and open question with some evidence supporting a strengthening of NL cultural identity under these conditions (Downes, 2001). This interpretation of our results is speculative and should be considered with caution as the strength of cultural identity was not assessed in the present study, preventing us from disentangling the separable influences of culture and language experience. Additional research is needed in order to better understand how these variables, either alone or in combination, modulate the MFLE.

Our exploratory analysis of the influence of FL experience on ratings of permissibility identified dilemma and dilemma type-specific findings. Unexpectedly, in more proficient English users, we observed that higher levels of FL dominance were associated with higher ratings of permissibility on the impersonal version of the *Burning Building* dilemma. Previous studies generally support a negative association between FL proficiency and the MFLE, manifesting as lower rates of utilitarian responding in more proficient bilinguals (Stankovic et al., 2022). The same association with FL proficiency has also been reported with respect to permissibility ratings where patterns of results generally mirror those observed in rates of utilitarian responding (e.g., Geipel et al., 2015a). Our finding, while unexpected in light of most previous reports, aligns with a recent result from Del Maschio et al. (2022b). In their investigation, bilinguals who were both highly proficient and dominant in their FL were more likely to judge the moral violation in the *Bike Week* dilemma as acceptable (i.e., killing another biker by kicking them off their motorcycle in order to avoid a larger accident with multiple fatalities). It should be noted that, unlike the *Bike Week* dilemma which was personal, other-beneficial, avoidable, and instrumental, the impersonal version of the *Burning Building* dilemma used in the present study was self-beneficial, avoidable, and accidental, suggesting that pattens of results are likely to differ based on the combination of characteristics for a specific dilemma, although this has yet to be explored. Additionally, on the personal version of the *Burning Building* dilemma, we again observed an unexpected finding: higher FL dominance associated with higher ratings of permissibility in less FL immersed bilinguals. Considering these results, and that our sample consisted entirely of bilinguals who learned and use English almost exclusively in a classroom environment, our findings may reflect reduced access to norms when processing dilemmas in an FL (Geipel et al., 2015a, 2015b). Due to the growing status of English as a language of education and business in China (Jin & Cortazzi, 2002), higher levels of proficiency and dominance may not correspond to a stronger connection between the FL and moral norms in Mandarin–English bilinguals. Finally, this counterintuitive pattern of findings was only observed in response to a self-beneficial dilemma, and that findings from the other-beneficial *Organ Transplant* dilemma revealed the negative association between FL proficiency and ratings of permissibility expected based on previous work. Additional work is needed in order to explore how separable dimensions of FL experience differentially influence responding to moral dilemmas that differ across the factors of Personal Force, Benefit Recipient, Evitability, and Intentionality.

We observed limited evidence in support of an influence of FL experience on ratings of distress. Based on the DECREASED EMOTIONALITY hypothesis (Costa et al., 2014), higher levels of FL experience, particularly FL proficiency, should be associated with higher ratings of distress due to increased activation of the emotional System I. This specific result was not observed on any dilemma. One possible explanation for our limited findings regarding ratings of distress is that the use of a brief introduction to contextualize dilemmas may have created additional distance between participants and the scenarios described, modulating the amount of distress experienced. While this is partially supported by the absence of a significant main effect of language condition on ratings of distress, this was not directly tested. Alternatively, the FL proficiency level of our sample may not have been high enough in order to observe the predicted reduction in emotionality when processing a dilemma in an FL. Surprisingly, higher FL proficiency associated with lower ratings of distress was observed on the personal version of the *Trolley* dilemma (i.e., *Footbridge* dilemma), but this effect was below our threshold for significance. Similar to results from models of ratings of permissibility, the only significant findings came in the form of interactions between separable dimensions of FL experience. Models of distress ratings identified a single significant interaction between FL immersion and FL dominance on the personal version of the *Trolley* dilemma. While the observed negative association between FL immersion and ratings of distress in less FL dominant bilinguals partially aligns with predictions generated from the DECREASED EMOTIONALITY hypothesis (i.e., higher FL experience associated with higher ratings of distress), the absence of significant main effects suggests that separable dimensions of FL experience do not have the same influence across all bilinguals on ratings of distress. The nature of these results aligns with the operationalization of bilingualism as a complex, multidimensional experience (Gullifer et al., 2021; Privitera et al., 2022b), and further underscores the importance of assessing dimensions beyond FL proficiency in future investigations.

In a recent meta-analysis, Stankovic et al. (2022) reported that the MFLE was modulated by self-reported FL proficiency with higher proficiency associated with lower levels of utilitarian responding. This finding is consistent with dual-system theories (Kahneman, 2011), and supports a unique role of FL proficiency in the modulation of the MFLE, especially in light of reported null

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press

influences of FL proficiency on the broader FLE (Del Maschio et al., 2022a). It is worth noting, however, that most previous studies of the MFLE utilize exclusively self-report assessments of FL proficiency, and generally forego the assessment of additional dimensions of FL experience. Our results, along with those of a recent investigation (Del Maschio et al., 2022b), highlight the importance of not only assessing but modeling differences across multiple dimensions of FL experience through the inclusion of both main effects and interaction terms, and call into question previously reported effects of separable dimensions of language experience investigated in isolation. However, further work is needed in order to better understand the nature of these interactions and how they modulate the MFLE.

Our reported findings should be considered in light of a few limitations. The dilemmas utilized in the present study do not represent an exhaustive set of moral situations, and results may have differed if an alternative set of dilemmas were used. Additionally, similar personal and impersonal version of these dilemmas were presented to participants in random order. While this practice has been used in previous studies to control for order effects (e.g., Christensen et al., 2014; Wong & Ng, 2018), and visual inspection of our data did not identify any obvious influence of presentation order, it is possible that participant responses were subtly influenced. Furthermore, the unique linguistic characteristics of our sample, native Mandarin speakers enrolled in an English-immersive university located in a Mandarin-immersive country, may limit generalization of our findings. Whether the linguistic environment or similarity between languages spoken influences the MFLE is an open question. Finally, while assessment of FL proficiency was objective, we relied exclusively on self-reported measures of FL immersion and FL dominance.

The current study investigated whether differences in separable dimensions of FL experience influenced the MFLE in a sample of Mandarin–English bilinguals. Using a set of both classic and contemporary sacrificial moral dilemmas, we identified evidence in support of a dilemma and dilemma-type specific MFLE. Further investigation of the modulating influence of FL experience identified interactions between separable dimensions on utilitarian responding and ratings of permissibility and distress. In total, these results support the conclusion that the influence of individual dimensions of FL experience on the MFLE is not the same across all bilinguals, and is instead modulated by differences in other dimensions. These findings emphasize the importance of considering differences in FL experience across multiple dimensions when investigating the MFLE. Future work should continue to explore the complex ways in which differences in separable dimensions of FL experience impact on the MFLE and other previously reported psycholinguistic phenomena.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/S1366728923000275

Acknowledgments. This work was supported by an internal research grant from Wenzhou-Kean University awarded to AJP [grant number WKUSSPF202206].

Competing interests. The authors declare none.

Data availability. The data that support the findings of this study, all code used in analyses, and complete model results are available on OSF (DOI 10.17605/OSF.IO/56YDZ) or from the corresponding author (aprivite@connect.hku.hk).

References

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*(2020), 10.

Anderson, J. A. E., Mak, L., Chahi, A. K., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, *50*(1), 250–263.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407.

Bereby-Meyer, Y., Hayakawa, S., Shalvi, S., Corey, J. D., Costa, A., & Keysar, B. (2020). Honesty speaks a second language. *Topics in Cognitive Science*, *12* (2), 632–643.

Białek, M., & Fugelsang, J. (2019). No evidence for decreased foreign language effect in highly proficient and acculturated bilinguals: A commentary on Čavar and Tytus (2018). *Journal of Multilingual and Multicultural Development*, *40*(8), 679–686.

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.

Black, L., & Fairbrother, H. (2008). The ethics of the elephant: why physician participation in executions remains unethical. *The American Journal of Bioethics*, *8*(10), 59–61.

Brouwer, S. (2019). The auditory foreign-language effect of moral decision making in highly proficient bilinguals. *Journal of Multilingual and Multicultural Development*, *40*(10), 865–878.

Caldwell-Harris, C. L. (2014). Emotionality differences between a native and foreign language: Theoretical implications. *Frontiers in Psychology*, *5*, 1055.

Caldwell-Harris, C. L. (2015). Emotionality differences between a native and foreign language: Implications for everyday life. *Current Directions in Psychological Science*, *24*(3), 214–219.

Campbell, T., O'Brien, E., Van Boven, L., Schwarz, N., & Ubel, P. (2014). Too much experience: a desensitization bias in emotional perspective taking. *Journal of Personality and Social Psychology*, *106*(2), 272.

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*(1), 17–28.

Cecchetto, C., Rumiati, R. I., & Parma, V. (2017). Promoting cross-culture research on moral decision-making with standardized, culturally-equivalent dilemmas: Th e 4CONFiDe set. *Journal of Health and Social Sciences*, *22*, 173–194.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264.

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.

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign-language effect. *Philosophical Psychology*, *29*(1), 23–40.

Circi, R., Gatti, D., Russo, V., & Vecchi, T. (2021). The foreign language effect on decision-making: A meta-analysis. *Psychonomic Bulletin & Review*, *28* (4), 1131–1141.

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*(7), 1109.

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.

Costa, A., Duñabeitia, J. A., & Keysar, B. (2019). Language context and decision-making: Challenges and advances. *Quarterly Journal of Experimental Psychology*, *72*(1), 1–2.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403.

De Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, *9*(3), 33.

Del Maschio, N., Crespi, F., Peressotti, F., Abutalebi, J., & Sulpizio, S. (2022a). Decision-making depends on language: A meta-analysis of the Foreign

Language Effect. *Bilingualism: Language and Cognition*, *25*(4), 617–630.

Del Maschio, N., Del Mauro, G., Bellini, C., Abutalebi, J., & Sulpizio, S. (2022b). Foreign to whom? Constraining the moral foreign language effect on bilinguals' language experience. *Language and Cognition*, 1–23.

Dewaele, J.-M. (2004). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development*, *25*(2-3), 204–222.

Downes, S. (2001). Sense of Japanese cultural identity within an English partial immersion programme: Should parents worry? *International Journal of Bilingual Education and Bilingualism*, *4*(3), 165–180.

Dylman, A. S., & Champoux-Larsson, M.-F. (2020). It's (not) all Greek to me: Boundaries of the foreign language effect. *Cognition*, *196*, 104148.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Geipel, J., Hadjichristidis, C., & Surian, L. (2015a). The foreign language effect on moral judgment: The role of emotions and norms. *PloS One*, *10*(7), e0131529.

Geipel, J., Hadjichristidis, C., & Surian, L. (2015b). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology*, *59*, 8–17.

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, *9*(1), 65–76.

Gollan, T. H., Salmon, D. P., Montoya, R. I., & Galasko, D. R. (2011). Degree of bilingualism predicts age of diagnosis of Alzheimer's disease in low-education but not in highly educated Hispanics. *Neuropsychologia*, *49*(14), 3826–3830.

Grant, A., & Li, P. (2019). Proficiency affects intra-and inter-regional patterns of language control in second language processing. *Language, Cognition and Neuroscience*, *34*(6), 787–802.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523.

Gullifer, J. W., Kousaie, S., Gilbert, A. C., Grant, A., Giroud, N., Coulter, K., Klein, D., Baum, S., Phillips, N., & Titone, D. (2021). Bilingual language experience as a multidimensional spectrum: Associations with objective and subjective language proficiency. *Applied Psycholinguistics*, *42*, 1–34.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21.

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*(10), 1387–1397.

Jin, L., & Cortazzi, M. (2002). English language teaching in China: A bridge to the future. *Asia Pacific Journal of Education*, *22*(2), 53–64.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

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.

Kinsella, C., & Singleton, D. (2014). Much more than age. *Applied Linguistics*, *35*(4), 441–462.

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*(5), 938–944.

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265–287.

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.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harper & Row.

Privitera, A. J., Momenian, M., & Weekes, B. S. (2022a). Graded bilingual effects on attentional network function in Chinese high school students. *Bilingualism: Language and Cognition*, *26*(3), 527–537. https://doi.org/10.1017/S1366728922000803

Privitera, A. J., Momenian, M., & Weekes, B. S. (2022b). Task-specific bilingual effects in Mandarin-English speaking high school students in China. *Current Research in Behavioral Sciences*, *3*.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Shin, H. I., & Kim, J. (2017). Foreign language effect and psychological distance. *Journal of Psycholinguistic Research*, *46*(6), 1339–1352.

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.

Stoica, P., & Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, *21*(4), 36–47.

Surrain, S., & Luk, G. (2017). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition*, *22*(2), 401–415.

Thomson, J. (1985). The trolley problem. *The Yale Law Journal*, *94*, 1395–1415.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Wermelinger, S., Gampe, A., & Daum, M. M. (2017). Bilingual toddlers have advanced abilities to repair communication failure. *Journal of Experimental Child Psychology*, *155*, 84–94.

Winskel, H., & Bhatt, D. (2020). The role of culture and language in moral decision-making. *Culture and Brain*, *8*(2), 207–225.

Wong, G., & Ng, B. C. (2018). Moral judgement in early bilinguals: Language dominance influences responses to moral dilemmas. *Frontiers in Psychology*, *9*, 1070.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125.

https://doi.org/10.1017/S1366728923000275 Published online by Cambridge University Press