Research Article

Lena Nadarevic*, Lena C. Klein, Janna Dierolf

Does foreign language alter moral judgments? Inconsistent results from two pre-registered studies with the CNI model

https://doi.org/10.1515/psych-2020-0112
received June 24, 2021; accepted December 1, 2021.

Abstract: Recent studies suggest that processing moral dilemmas in a foreign language instead of the native language increases the likelihood of moral judgments in line with the utilitarian principle. The goal of our research was to investigate the replicability and robustness of this moral foreign-language effect and to explore its underlying mechanisms by means of the CNI model—a multinomial model that allows to estimate the extent to which moral judgments are driven by people's sensitivity to consequences (*C*-parameter), their sensitivity to norms (*N*-parameter), and their general preference for action or inaction (*I*-parameter). In two pre-registered studies, German participants provided moral judgments to dilemmas that were either presented in German or English. In Experiment 1, participants judged eight different dilemmas in four versions each (i.e., 32 dilemmas in total). In Experiment 2, participants judged four different dilemmas in one of the four versions (i.e., 4 dilemmas in total). Neither of the two studies replicated the moral foreign-language effect. Moreover, we also did not find reliable language effects on the three parameters of the CNI model. We conclude that if there is a moral foreign-language effect, it must be quite small and/or very fragile and context specific.

Keywords: moral judgment; moral dilemmas; foreign-language effect; CNI model; multinomial modeling.

1 Introduction

Imagine the following situation, taken from Gawronski et al. (2017):

> *You are the director of a hospital in a developing country. A foreign student who is volunteering in the country got infected with a rare virus. The virus is highly contagious and deadly to seniors and children. The student suffers from a chronic immune deficiency that will make her die from the virus if she is not returned to her home country for special treatment. However, taking her out of quarantine involves a considerable risk that the virus will spread. Would you take the student out of quarantine to return her to her home country for treatment in this case?*

With the outbreak of the corona virus disease 2019 (Covid-19) similar dilemmas as the one described above became sad reality for decision-makers throughout the world. For example, after the first diagnosed cases of Covid-19 in Wuhan, China, governments had to decide whether they should fly their citizens out of Wuhan despite the risk of bringing Covid-19 to their homelands. A choice on this matter represents a moral dilemma because different ethical principles are in conflict to each other. In the following, we will address the broad distinction between *deontological principles* and the *utilitarian principle*.

***Corresponding author: Lena Nadarevic,** Department of Psychology, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany, E-mail: nadarevic@psychologie.uni-mannheim.de, ORCID: 0000-0003-1852-5019
Lena C. Klein, Janna Dierolf, Department of Psychology, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany

Judging an action based on deontological principles means that the moral acceptability of the action is evaluated primarily by its conformity to such principles (e.g., given moral norms or obligations) rather than by its (expected) consequences. For example, from a deontological perspective one might argue that it is a government's moral duty to take responsibility for citizens in need which also entails evacuating citizens from the epicenter of a virus outbreak. In contrast, judging an action based on the utilitarian principle means that the action is evaluated by its (expected) consequences in the first place. An action is considered moral acceptable if, compared to any other option, it maximizes the net benefit for the greater good (e.g., in terms of lives saved relative to lives lost). For example, from a utilitarian perspective one might argue that it is inappropriate to evacuate citizens from the epicenter of a virus outbreak because doing so risks spreading the virus and threatening the entire population. In view of the dilemma described above, the German government decided to evacuate its citizens and their relatives from Wuhan, thus opting for the deontological option.[1] But would German politicians have made the same choice if they had discussed the dilemma in English rather than in their mother tongue (e.g., in a European committee)?

At first sight, this question seems somewhat bizarre. However, there is evidence for a *moral foreign-language effect* (MFLE, Hayakawa et al., 2017) in the literature. The MFLE is characterized by a higher proportion of utilitarian choices when moral dilemmas are presented in a foreign language than when they are presented in the mother tongue (Cipolletti et al., 2016; Corey et al., 2017; Costa et al., 2014; Geipel et al., 2015a). Importantly, the effect has been found for a variety of native and foreign languages (Cipolletti et al., 2016a; Corey et al., 2017b; Costa et al., 2014b; Geipel et al., 2015a). Therefore, it is rather unlikely that the effect relies on different cultural norms triggered by the different languages. Interestingly, the MFLE has been primarily reported for personal moral dilemmas (Costa et al., 2014; Geipel et al., 2015a), i.e., dilemmas in which the utilitarian choice requires harming or killing one or more individuals by personal force (Moore et al., 2011). The prime example of a personal dilemma is the *footbridge dilemma* (Thomson, 1976) in which a runaway trolley is about to kill five people on the tracks. The only way to save the five is to push a large bystander from a footbridge in front of the trolley to stop it. The impersonal equivalent of this trolley scenario is the *switch dilemma* (Foot, 1967) in which the only way to save the five people is to divert the trolley on a side track where it will kill a single person instead of the five people on the main track. A mini-meta analysis by Geipel et al. (2015a) revealed a MFLE of size d = 0.52, 95% CI [0.34, 0.71] for the footbridge dilemma. For the switch dilemma, in contrast, the effect size was considerably smaller with confidence intervals including zero, d = 0.11, 95% CI [-0.07, 0.29]. Likewise, later studies replicated the MFLE for the footbridge dilemma, but not for the switch dilemma (Chan et al., 2016; Cipolletti et al., 2016; Corey et al., 2017, but see Experiment 1).

According to the dual-process model of Greene et al. (2001), personal dilemmas (e.g., the footbridge dilemma) elicit much stronger emotional responses than impersonal ones (e.g., the switch dilemma). More precisely, the use of personal force, inherent in the utilitarian response option, triggers a strong negative emotional response. This leads to an intuitive preference for the deontological option. However, a controlled calculation of costs and benefits can override this emotional response leading to a preference for the utilitarian option. Based on this dual-process model, Costa et al. (2014) reasoned that the MFLE is due to a reduced emotionality in a foreign language which in turn increases the likelihood of a controlled cost-benefit analysis favoring the utilitarian option. Indeed, there is psychophysiological evidence that people respond less emotionally to foreign-language stimuli (e.g., Harris, 2004; Thoma & Baum, 2019). However, Geipel et al. (2015a, Study 2) did not find the MFLE to be mediated by different levels of reported emotional distress. Moreover, in one of their experiments the authors replicated the MFLE for a low-emotion, impersonal dilemma (lost wallet dilemma), but not for a high-emotion, personal dilemma (crying baby dilemma). This finding is difficult to reconcile with Costa et al.'s *emotional-distancing hypothesis*. For this reason, Geipel and colleagues introduced an alternative explanation of the MFLE. They argued that social and moral norms are overall less accessible in a foreign language (*norm-accessibility hypothesis*). In a subsequent article (Geipel et al., 2015b), the authors came up with yet another explanation of the MFLE. They proposed that foreign language leads to a different weighting of intentions and consequences with a stronger focus on consequences as compared to the native language (*weighting hypothesis*).

To reiterate, according to the norm-accessibility hypothesis, foreign language should reduce norm-based choices, i.e., choices based on deontological considerations. In contrast, according to the weighting hypothesis, foreign language should increase consequence-based choices, i.e., choices based on utilitarian considerations. Unfortunately,

1 For simplicity, we refer to the dilemma option favored on the basis of deontological principles as the deontological option and the dilemma option favored by the utilitarian principle as the utilitarian option.

the two predictions are not testable with standard moral dilemmas in which a decision for the deontological option implies the rejection of the utilitarian option, and vice versa. To address this confound present in standard moral dilemmas, Conway and Gawronski (2013) developed a process dissociation procedure (PDP) that allows to disentangle and to measure people's deontological and utilitarian inclinations. This procedure requires presenting participants with several dilemmas in two different versions each. The so-called *incongruent version* is the standard dilemma version as introduced above (e.g., the classical footbridge dilemma). Here the behavior at choice is acceptable from the utilitarian perspective but unacceptable from the deontological perspective (e.g., killing an innocent person in order to save five). In contrast, in the so-called *congruent version*, the behavior at choice is unacceptable from the deontological *and* the utilitarian perspective (e.g., killing an innocent person in order to save another one). Hence, a choice based on deontological considerations should always lead to an "unacceptable" judgment, irrespective of the dilemma version. In contrast, a choice based on utilitarian consideration should lead to an "unacceptable" judgment for congruent dilemmas but not for incongruent ones. The observed differences in the proportion of "unacceptable" judgments between the two dilemma versions thus reflect a person's utilitarian inclination. A person's deontological inclination, in contrast, is determined by the proportion of "unacceptable" judgments for the incongruent dilemmas divided by the proportion of "unacceptable" judgments for the congruent dilemmas that cannot be explained by utilitarian inclinations.

Muda et al. (2017) were the first who used Conway and Gawronski's (2013) PDP to explore the underlying mechanisms of the MFLE. In line with previous studies, the authors compared moral dilemma choices between language groups (native language group: Polish; foreign-language group: English) for the incongruent dilemmas in a first step. With this traditional analysis, the authors did not find the MFLE. Yet subsequent data analyses with the PDP revealed lower deontological inclinations in the foreign-language group compared to the native-language group. Surprisingly, however, the same was true for participants' utilitarian inclinations. A later PDP study including other native and foreign languages and different question formats led to similar results (Hayakawa et al., 2017). Taken together, these findings indicate that people are less moral (i.e., less deontological *and* less utilitarian) in a foreign language. However, because the MFLE did not replicate in these studies, no conclusions can be drawn about its underlying mechanisms. Moreover, despite the advantages of the PDP over the traditional dilemma approach, both methods still share the following problem: they confound moral principles with general (in)action preferences (Crone & Laham, 2017; Gawronski et al., 2016). In the incongruent dilemmas the utilitarian option is always linked to an action (e.g., pushing a bystander from a footbridge) whereas the deontological option is always linked to inaction (e.g., not doing anything). Hence, if the action is judged as "acceptable", this judgment can reflect utilitarian inclinations or a general preference for action. Likewise, an "unacceptable" judgment can be based on deontological inclinations or a general preference for inaction. Accordingly, the MFLE could also be due to an increased action preference (or decreased inaction preference, respectively) in a foreign language.

To disentangle moral principles from general response preferences, Gawronski et al. (2017) introduced a multinomial processing tree model (see Erdfelder et al., 2009, for a review) tailored to moral dilemma judgments. This so-called *CNI* model is named after its three parameters, each of which represents a different component underlying moral judgments. The *C*-parameter represents people's sensitivity to consequences (i.e., their utilitarian inclinations), the *N*-parameter represents people's sensitivity to norms (i.e., their deontological inclinations), and the *I*-parameter reflects people's preference for inaction (or action, respectively). In order to estimate these parameters, the model requires judgments on four different versions of a moral dilemma (V1 to V4), which are defined by the following characteristics: V1) The described action violates a moral norm but the action benefits exceed the costs, V2) the action violates a moral norm and the action costs exceed the benefits, V3) the actions satisfies a moral norm and the action benefits exceed the costs, V4) the action satisfies a moral norm, but the action costs exceed the benefits. The CNI model's predictions are illustrated in Figure 1.

The model denotes that with probability *C*, moral judgments rely on people's sensitivity to consequences. This means that only actions with a net benefit are judged as acceptable (as in V1 and V3). If sensitivity to consequences does not underlie moral judgments (probability 1-*C*), people's sensitivity to norms drives these judgements with probability *N*. This means that only actions satisfying prescriptive moral norms are judged as acceptable (as in V3 and V4). Finally, if moral judgments are not based on a sensitivity to norms either (probability 1-*N*), the judgments rely on general (in)action preferences. In case of a preference for inaction over action (probability *I*), the described actions will be judged as unacceptable. In contrast, in case of a preference for action over inaction (probability 1-*I*), they will

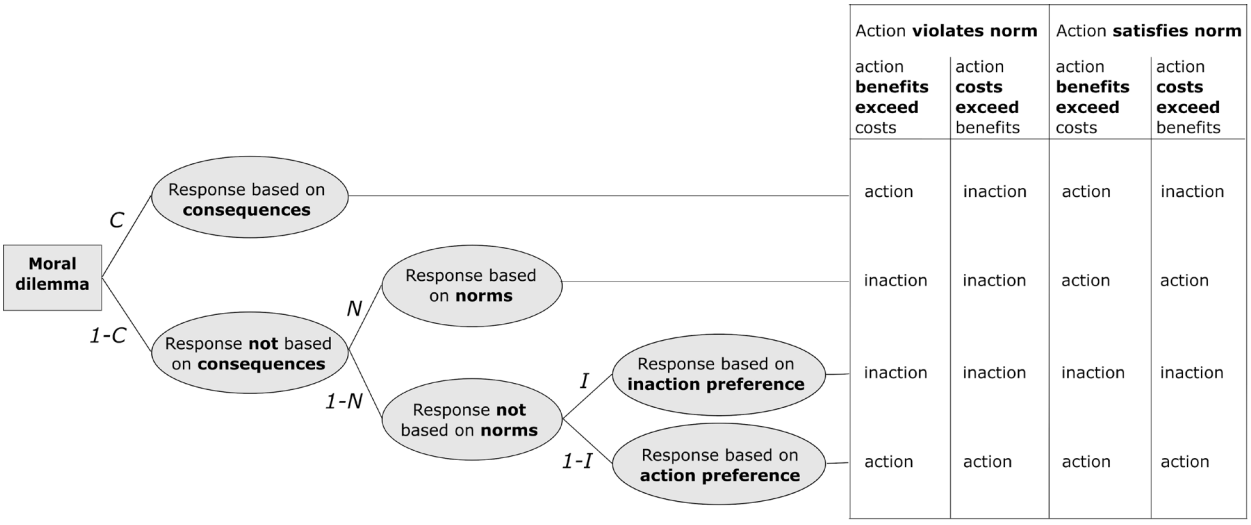

Figure 1: The CNI model of moral decision making.
Note. Graphical illustration of the CNI model by Gawronski et al. (2017). The processing tree on the left-hand side denotes different events that can underlie an observable response (i.e., action or inaction, respectively). The parameters attached to the branches of the tree reflect the event probabilities. The table on the right-hand side displays each predicted response, depending on the processing event (rows) and the dilemma version (columns).

be judged as acceptable. The absolute frequencies with which the actions in the different dilemma versions are judged as appropriate or inappropriate allow to estimate the parameters of the CNI model by multinomial modelling. For background information and a practical guide on multinomial processing tree modeling see, for example, Moshagen (2010) or Singmann and Kellen (2013).

The goal of this work was to use the CNI model to investigate foreign-language effects on moral dilemma judgments. More precisely, we aimed to test the dilemma specificity of the MFLE and to find out whether the MFLE is based on a) a reduced sensitivity to moral norms in the foreign language, b) an increased sensitivity to consequences in the foreign language, c) a reduced preference for inaction in the foreign language, or d) any combination of these three mechanisms. The two studies conducted were pre-registered and all data and materials are publicly available on the Open Science Framework (OSF; https://osf.io/b2as9/). For both studies, we will report how we determined our samples sizes, all data exclusions, all manipulations, and all measures.

2 Experiment 1

In Experiment 1, we compared moral dilemma judgments across two groups of participants who were native speakers of German and had learned English as a foreign language. In the native-language group we presented all instructions and materials in German, whereas in the foreign-language group we presented them in English. Participants of both groups judged eight different dilemmas in four versions each (i.e., 32 dilemmas in total). Six of the eight dilemmas were taken from Gawronski et al. (2017). These dilemmas had been used by Gawronski and colleagues to validate their CNI model. However, at the time we planned our experiments, we were not aware of any MFLE studies that had used these dilemmas. For this reason, we chose to add the footbridge dilemma and the terrorist dilemma which had successfully produced a MFLE in previous studies (e.g., Corey et al., 2017).

Simultaneously with our project, Białek et al. (2019) also explored the MFLE with the CNI model. In their study, Polish participants either judged moral dilemmas in their native tongue or in one of four foreign languages (English, German, Spanish, or French). The dilemmas presented were the ones developed by Gawronski et al. (2017) for their CNI model. However, Białek et al. (2019) did not find the MFLE with these dilemmas. Moreover, their model-based analyses revealed findings similar to the ones of previous studies with the PDP (Hayakawa et al., 2017; Muda et al., 2017). That is, foreign

language reduced participants' sensitivity to norms and their sensitivity to consequences. What is noticeable about the above-mentioned studies is that the dilemmas presented were mainly impersonal ones, i.e., dilemmas that do not involve personal force. Thus, possibly these studies could not replicate the MFLE because the effect primarily occurs with personal dilemmas. Moreover, the studies did not involve an objective measure of foreign-language proficiency. Research suggests that individuals who are highly proficient in the foreign language do not show the MFLE (Čavar & Tytus, 2018, but note that a control group was missing in this study). However, the fact that sensitivity to norms as well as sensitivity to consequences were lower in the foreign-language compared to the native-language conditions rather speaks against this explanation. In fact, the opposite may have been the case. Greater difficulty in understanding the dilemmas in the foreign language may have led to more random answers and thus "less moral" answers. In our experiments, we chose to not only assess language proficiency with self-report measures, but also with the LexTALE (Lemhöfer & Broersma, 2012), an objective measure of language proficiency. Moreover, following Chan et al. (2016), we implemented a comprehension test for the dilemmas to identify participants who did not understand the dilemma contents.

We preregistered Experiment 1 and its hypotheses on March, 5, 2018 with the template from AsPredicted.org on the Open Science Framework (https://osf.io/3e4mz). We predicted that we would replicate the MFLE with the footbridge and the terrorist dilemma (Hypothesis 1a). Because at the time of preregistration the study by Białek et al. (2019) had not been published yet, we were also optimistic to find the effect with Gawronski et al.'s (2017) dilemma set (Hypothesis 1b). In addition, we set up the following predictions for the CNI model analyses. If the MFLE relies on a reduced sensitivity to moral norms in the foreign language as compared to the native language, the CNI model's *N*-parameter will be reduced in the foreign-language condition (Hypothesis 2a). In contrast, if the MFLE relies on an increased sensitivity to consequences in the foreign language as compared to the native language, the *C*-parameter will be increased in the foreign-language condition (Hypothesis 2b). Finally, if the MFLE relies on a reduced preference for inaction in the foreign language as compared to the native language, the *I*-parameter will be reduced in the foreign language condition (Hypothesis 2c). Note that hypotheses 2a, 2b, and 2c are not mutually exclusive, as several mechanisms may underlie the MFLE.

2.1 Method

2.1.1 Participants

We aimed for a target sample size of 80 participants. A power-analysis with G*Power (Faul et al., 2009) indicated that with this sample size, a χ^2-test (df = 1, α = .05) would have a power of .80 to detect a medium size MFLE of $w \geq .31$ for a single dilemma and a small MFLE of $w \geq .11$ for an aggregated analysis of the eight dilemmas in their standard form (V1). We did not perform a power analysis for the CNI-model analyses. This is because at the time we planned Experiment 1, we were not aware of any CNI-model study on the MFLE from which we could have used the values as benchmark for this analysis.

Participants were recruited from March 12 to 16, 2018 at the University of Mannheim, Germany. A total of 79 participants completed the experiment. However, in line with our pre-registered exclusion criteria, we excluded 15 participants for the following reasons (two of which applied concurrently to one participant): non-native German speaker (n = 1), declaration of non-serious participation (n = 1), accuracy below 50% in the comprehension test (n = 3). Moreover, as the MFLE seems to be smaller for individuals that are highly proficient in the foreign language (e.g., Corey et al., 2017), we followed Costa et al. (2014) and excluded participants who had spent more than ten months in an English-speaking country (n = 11). We had also preregistered to exclude participants who rated their English proficiency as mother tongue or who indicated to have a parent whose native language was English. However, the latter two criteria did not apply to any participant. Thus, the final sample comprised 64 native German speakers (51 female, 13 male) most of which were psychology students (92%). The age of the participants ranged from 18 to 28 years (M = 22.0, SD = 2.9). Demographic information by condition is provided in the Appendix (see Table A1).[2]

2 Six participants of the final sample failed a one-item attention check. We did not exclude these participants, because a) this had not been specified in the pre-registration, and b) the main pattern of results remained the same with and without these participants.

Table 1: Summary of the moral dilemmas presented in Experiment 1.

Dilemma	Summary
Abduction	As your country's president, you have to decide whether to veto the payment of a ransom of one million dollars for a kidnapped journalist to a guerilla group because this money will eventually be used to kill many other people.
Transplant	As the surgeon of a small hospital, you have to decide whether to terminate the life support of a patient who is unlikely to wake up again in order to use his organs to save five people injured in a car accident who would otherwise die from their injuries.
Torture	As a member of a special police department, you have to decide whether to use illegal interrogation techniques such as torture on a man accused of having kidnapped several children in order to find the kidnapped children in time and save them from dehydration.
Assisted Suicide	As a doctor, you have to decide whether to give a fatally ill patient the lethal medicine he has been asking for in order to end his terrible pain.
Immune Deficiency	As the director of a hospital in a developing country, you have to decide whether to give medicine to a foreign student who caught a rare virus that is highly contagious and deadly to seniors and children. Due to a chronic immune deficiency, the student would die from the side effects of the medicine, but not die from the virus.
Vaccine	As a doctor in an area that suffers from an outbreak of a highly contagious virus, you have to decide whether to use a vaccine that would help preventing the spread of the virus but kill several people due to its severe side-effects.
Footbridge	As a bystander on a footbridge, you have to decide whether to push a large, strange man off the bridge in front of an out-of-control train, causing the man's death, in order to prevent the train from killing five workmen on the tracks.
Terrorist	As a negotiator you have to decide between the following options dictated by a group of terrorists who have captured a group of six tourists. Either you yourself shoot one of the tourists and the other five will be released, or the terrorists will release one tourist and kill the other five.

Note. All summaries of the dilemmas refer to the standard dilemma version (V1)

2.1.2 Materials

We used eight different dilemmas in four versions each. Six of the eight dilemmas were taken from Gawronski et al. (2017), who had listed their dilemmas in English and German in the supplementary material of their article. We slightly modified the wording of these CNI-model dilemmas to ensure perfect comparability of the dilemmas between languages. The dilemmas captured the following topics: abduction, transplant, torture, euthanasia, immune deficiency, and vaccine. In addition, we used the footbridge dilemma (Thomson, 1976) and the terrorist dilemma (Corey et al., 2017, adapted from Greene et al., 2001). For both personal dilemmas, previous studies had found the MFLE (e.g., Corey et al., 2017; Geipel et al., 2015b; but see Muda et al., 2020).

We complemented the standard version of these dilemmas (V1: action violates norm, action benefits exceed action costs) by three additional versions that are required for a CNI model analysis (V2: action violates norm, action benefits do not exceed costs, V3: action satisfies norm, action benefits exceed action costs, V4: action satisfies norm, action benefits do not exceed action costs). One of the authors (LK) translated the dilemmas to German. The translations were then cross-checked by another author (LN). Finally, we assigned all moral dilemmas to four different sets. Each set contained eight thematically different dilemmas. Per set, each type of dilemma version (i.e., V1, V2, V3, or V4) was represented twice. Table 1 provides a short description of the standard version (i.e., V1) of each dilemma.

To check whether all participants had truly understood the presented dilemmas, we constructed a comprehension test. Similar to Chan et al. (2016), this test consisted of several multiple-choice questions, one question for each response alternative of a dilemma. To limit the length of the comprehension test, the questions only referred to the standard version (V1) of a dilemma. The comprehension test thus contained 16 multiple choice questions in total, i.e., two questions for each of the eight thematically different dilemmas. The multiple-choice questions involved four

Table 2: Comprehension test questions for the footbridge dilemma.

Question	Answer options
1. What will happen if you push the stranger onto the tracks?	a. The five workmen will die b. The five workmen will live c. The stranger will die d. The stranger will live
2. What will happen if you don't push the stranger onto the tracks?	a. The five workmen will die b. The five workmen will live c. The stranger will die d. The stranger will live

response options of which either one or two were correct (see Table 2 for an example). All dilemmas and comprehension test questions are available at: https://osf.io/b2as9/.

In order to measure participants' English proficiency in the foreign-language condition, we used the *Lexical Test for Advanced Learners of English* (LexTALE, Lemhöfer & Broersma, 2012). The LexTALE is a lexical-decision task that consists of 40 words and 20 non-words. For each item, participants must decide whether it is an existing English word. The items are presented in a fixed order and the test does not involve a time constraint. We also used the German version of the LexTALE to measure participants' German proficiency in the native-language condition. According to Lemhöfer and Broersma (2012), LexTALE scores are comparable between the English and German test, thus allowing us to compare language proficiency between the two language conditions.

2.1.3 Design

The research design was a 2 (group: foreign language, native language) × 4 (dilemma version: V1, V2, V3, V4) mixed design with group manipulated between participants and dilemma version manipulated within participants. Moreover, four dilemma sets were counterbalanced across four experimental blocks by means of a Latin square. The dependent variable was the moral judgment that participants provided in response to each dilemma. Participants judged a described action as either "appropriate" or "inappropriate".

2.1.4 Procedure

The study took place in the laboratory. After providing informed consent, participants were randomly assigned to the native-language or foreign-language group. Depending on the group, the computer presented all instructions and materials in German or English. All participants were informed that they would see several short stories, some of which would appear similar, but differ in important ways. For this reason, participants were instructed to read all contents very carefully. Moreover, participants were told that their task was to judge the appropriateness of different actions described by the stories. Participants then judged 32 moral dilemmas that were presented in four sets assigned to four subsequent blocks. Before each new block, they were again reminded to read all contents very carefully. Moreover, participants had to take a short break of at least 30-sec after each block to stay focused. Each dilemma appeared on the screen until participants judged the moral appropriateness of the described action ("*yes, I find the action appropriate*" vs. "*no, I find the action inappropriate*").

Following the moral judgment phase, a lengthy text appeared on the screen that included an instruction to answer "very bad" on a mood item displayed below. In line with Gawronski et al. (2017), this item served as an attention check. Participants then answered the 16 comprehension test questions. Because this test was designed to check participants' reading comprehension and not their memory, the computer presented all V1-dilemmas for a second time. Each of these dilemmas appeared on the top of the screen together with two comprehension questions displayed below. Following the comprehension test, the foreign-language group evaluated their overall English proficiency (*beginner,*

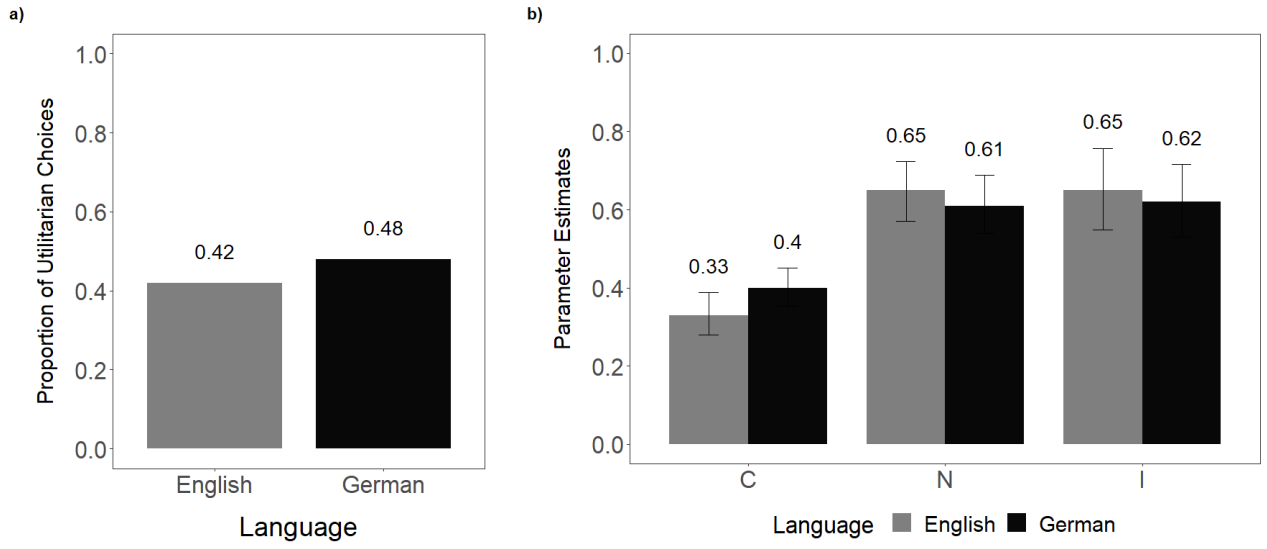

Figure 2: Results of a) the conventional analysis and b) the CNI analysis for the foreign-language group (English) and the native-language group (German) of Experiment 1.

basic, conversational, fluent, close to native, mother tongue) and rated their English skills on four different dimensions (oral production, written production, oral comprehension, and written comprehension) on a scale ranging from *very poor* (1) to *excellent* (7). Participants in this group also indicated whether they held an English certificate, whether one parent was a native English speaker, when they had started to learn English, and how many months they had spent in an English-speaking country (see Table A2 in the Appendix). They were also asked to indicate the number of moral dilemmas from the judgment phase they had difficulties understanding. Next, participants in both groups completed the LexTALE. The foreign-language group received the English test, and the native-language group received the German test. Finally, all participants were asked whether they had answered all questions seriously and whether they had been distracted at any point during the experiment.

2.2 Results

We had pre-registered the following analyses: 1) conventional dilemma analyses with χ^2-tests, 2) analyses with the PDP approach, 3) multinomial CNI model analyses, and 4) secondary analyses testing the robustness of our findings.

2.2.1 Conventional analyses

To test for an overall MFLE across dilemmas, we analyzed the dilemma choices for the V1-dilemmas only. The V1-dilemmas are characterized by a proscriptive norm and action benefits larger than action costs. This corresponds to the standard dilemma version that has been used by most prior MFLE studies (e.g., Costa et al., 2014; Geipel et al., 2015b). In a first step, we computed the proportion of "utilitarian choices" across all V1-dilemmas separately for the two experimental groups. Surprisingly, this proportion of utilitarian choices was slightly higher in the native-language group than in the foreign-language group (see Figure 2a), albeit not significantly, $\chi^2(1) = 1.80$, $p = .180$, $\varphi = .06$, 95% CI [.00, .15]. In a second step, we repeated this analysis on the item level. However, again, we did not find the predicted MFLE for any of the eight dilemmas, χ^2s(1) ≤ 2.77, *p*s ≥ .096, φs ≤.21. Moreover, only one dilemma, the torture dilemma, showed a descriptive pattern in the expected direction (see Table 3). Hence, unlike predicted by Hypotheses 1a and 1b, we neither replicated the MFLE for the footbridge dilemma or the terrorist dilemma nor for any of the six CNI-model dilemmas.

Table 3: Proportion of utilitarian choices for the standard version (V1) of each dilemma as a function of language (English = foreign language; German = native language) in Experiment 1.

Dilemma	English	German	χ^2-test (df = 1)	φ with .95 CI
Abduction	.61	.64	$\chi^2 = 0.07$, $p = .795$	.03 [.00, .26]
Transplant	.25	.28	$\chi^2 = 0.06$, $p = .803$	.03 [.00, .26]
Torture	.71	.56	$\chi^2 = 1.69$, $p = .193$	.16 [.00, .41]
Assisted Suicide	.75	.83	$\chi^2 = 0.68$, $p = .411$	.10 [.00, .35]
Immune Deficiency	.14	.28	$\chi^2 = 1.68$, $p = .195$	.16 [.00, .41]
Vaccine	.54	.61	$\chi^2 = 0.37$, $p = .545$	.08 [.00, .32]
Footbridge	.04	.17	$\chi^2 = 2.77$, $p = .096$	.21 [.00, .45]
Terrorist	.32	.47	$\chi^2 = 1.48$, $p = .223$	.15 [.00, .40]

2.2.2 Process dissociation analyses

Next, we calculated participants' utilitarian inclinations (*U*) and their deontological inclinations (*D*) based on the PDP approach suggested by Conway and Gawronski (2013):

(1) $U = p(\text{inappropriate/congruent}) - p(\text{inappropriate/incongruent})$
(2) $D = p(\text{inappropriate/incongruent})/(1\text{-}U)$

In line with Gawronski et al. (2017) we only included the responses for the proscriptive, incongruent dilemmas (V1 dilemmas) and the proscriptive, congruent dilemmas (V2 dilemmas) to calculate *U* and *D*. That is, responses to the prescriptive dilemmas (V3 and V4) were not considered in the following analyses. Additionally, we excluded one participant who had a *U*-parameter of 1, because it is not possible to calculate *D* in this case. Because *U* and *D* have different ranges, we *z*-standardized both parameters for the following analyses (see Conway & Gawronski, 2013). Unlike Muda et al. (2017), we did not find a significant difference in *U* between the native-language group ($M = 0.06$, $SD = 1.06$) and the foreign-language group ($M = -0.08$, $SD = 0.93$), $t(61) = 0.56$, $p = .574$, $d = 0.14$, 95% CI [-0.35, 0.64]. Moreover, there was also no significant difference in *D* between groups, $t(61) = 1.22$, $p = .226$, $d = 0.31$, 95% CI [-0.19, 0.81]. However, on the descriptive level, *D* was higher in the foreign-language group ($M = 0.17$, $SD = 0.73$) than in the native-language group ($M = -0.14$, $SD = 1.16$), which is at odds with the pattern of results obtained by Muda and colleagues as well as with the norm-accessibility hypothesis of the MFLE.

2.2.3 CNI model analyses

Finally, we analyzed moral judgments with the CNI model to test whether language influenced participants' sensitivity to norms (Hypothesis 2a), sensitivity to consequences (Hypothesis 2b), and their general preference for inaction versus action (Hypothesis 2c). This analysis included all moral judgments, i.e., judgments on all four versions of each dilemma. The model for the aggregate data fit the data well, $G^2(2) = 0.38$, $p = .829$. Parameter estimates are displayed in Figure 2b. Descriptively, participants' sensitivity to consequences was smaller in the foreign language compared to the native language. However, this difference was not statistically significant, $\Delta G^2(1) = 3.43$, $p = .064$. Likewise, language neither affected participants' sensitivity to norms, $\Delta G^2(1) = 0.36$, $p = .547$, nor their general preference for inaction versus action, $\Delta G^2(1) = 0.18$, $p = .674$. Both language groups showed a preference for inaction as indicated by *I*-parameters significantly larger than .50, ΔG^2s$(1) \geq 6.43$, $p \leq .011$.

We also fit the CNI model separately to each of the eight different dilemmas. The model provided a good fit to all data sets, G^2s$(2) \leq 4.33$, $p \geq .115$. However, as parameter estimates were based on a smaller amount of data when fitting

the model per dilemma, the estimates were quite unreliable as indicated by relatively large confidence intervals (see Table A3 in the Appendix for parameter estimates per dilemma). Thus, it is not surprising that, with few exceptions, parameters did not differ significantly between language groups. The only significant differences were the following: The *C*-parameter was significantly smaller in the foreign-language group compared to the native language group in the abduction dilemma, $\Delta G^2(1) = 9.75$, $p = .002$, as well as in the footbridge dilemma $\Delta G^2(1) = 4.20$, $p = .040$. Regarding the *N*-parameter, we again observed significant differences in the abduction dilemma, with a smaller parameter estimate in the foreign-language group compared to the native language group, $\Delta G^2(1) = 4.66$, $p = .031$. In contrast, the reverse was true for the transplant dilemma for which the *N*-parameter was significantly larger in the foreign-language group compared to the native-language group, $\Delta G^2(1) = 5.23$, $p = .022$. Finally, there were no significant group differences in the *I*-parameter.

Note that based on the multiple parameter comparisons made, the reported significant differences on the item level should be interpreted with caution. This is especially the case as Gawronski et al. (2020) recently expressed concerns about the construct validity of the abduction dilemma, which accounted for two of the four reported significant effects that we observed on the item level. More relevant than statistical significance, however, is that we found large variations of the data pattern across dilemmas, both in terms of the magnitude of the three model parameters and in terms of the direction of descriptive group differences. These differences suggest that it is reasonable to estimate parameters separately for individual dilemmas instead of estimating parameters based on the aggregate data. In fact, parameter heterogeneity violates an underlying assumption of MPT modeling and thus can lead to biased parameter estimates as well as biased statistical inferences (e.g., Klauer, 2006; Matzke et al., 2015).

2.2.4 Secondary analyses

In the pre-registration we had also indicated to compare the data of a comprehension test between language groups to make sure that participants in the foreign-language group did not experience more difficulties in understanding the moral dilemmas than participants in the native-language group. The number of correctly answered comprehension-test questions was relatively high and very similar between the native-language group ($M = 13.7$, $SD = 1.9$) and the foreign-language group ($M = 13.1$, $SD = 2.4$), $t(62) = 1.16$, $p = .250$, $d = 0.29$, 95% CI [-0.21, 0.79]. In addition, we had specified that if we do not find the MFLE for any dilemma, we will have a closer look at participants' language proficiency as measured by the LexTALE. Doing so is important to rule out the possibility that participants were equally proficient in German and English. However, this was not the case. Language proficiency was significantly higher in the native language (German LexTALE score: $M = 89.8$, $SD = 3.8$) than in the foreign language (English LexTALE score: $M = 72.5$, $SD = 9.7$), $t(33.59) = 8.92$, $p < .001$, $d = 2.35$, 95% CI [1.51, 2.97]. The absence of the MFLE can thus not be attributed to equivalent language proficiencies. Finally, we tested the robustness of our findings when excluding six further participants (three in each language group) who had failed the one-item attention check. However, even with this more stringent criterion, the data pattern and significance level of our analyses remained essentially the same. In fact, group differences were even smaller on the descriptive level. A detailed description of the results is provided on the OSF (https://osf.io/w8zma/).

2.3 Discussion

Experiment 1 failed to replicate the MFLE, both when analyzing moral judgments across dilemmas and when analyzing the judgments separately for each dilemma. This was the case even for the footbridge dilemma and the terrorist dilemma, two personal dilemmas for which the effect had previously been reported. What is more, we also did not find systematic differences in the CNI model parameters between language groups. When fitting the model per dilemma, parameter estimates varied drastically between the eight presented dilemmas. Yet for the few dilemmas for which we observed significant parameter differences between language groups the data pattern was inconsistent.

One criticism that can be raised against Experiment 1 is that the different dilemma versions varied within participants, in line with other CNI model studies. This resulted in a large number of moral dilemmas per participant (viz. 32 dilemmas in total) which may have fostered inattention to important details of the dilemmas. Moreover, there is empirical evidence that moral judgments are susceptible to sequence effects. For example, Wiegmann et al. (2012)

observed that when the footbridge dilemma preceded the switch dilemma, the action in the latter was considered less appropriate than usual. Furthermore, the authors showed that sequence effects were stronger for similar dilemmas. Although it is unclear whether and to what extent sequence effects contributed to the findings of Experiment 1, it is striking that the proposed MFLE also did not occur in other studies that manipulated different versions of the same dilemma within subjects (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2017).

What is also critical about Experiment 1 is that the final sample size (n = 64) was smaller than the target size (n = 80). In fact, a post-hoc power analysis indicated a power of .70 (instead of .80) to detect a MFLE of $w \geq .31$ when analyzing the data of a single dilemma. This relatively low power for the dilemma-wise analysis, however, cannot explain why only one out of eight dilemmas showed a data pattern in the expected direction. What is more, even when relaxing the exclusion criteria so that our analyses comprised a larger sample, we did not find a MFLE for any dilemma. More precisely, as the exclusion of participants who had spent more than ten months in an English-speaking country can be criticized for being arbitrary, we repeated our conventional analyses without this exclusion. This increased the sample size to n = 74. The only notable difference in results was a significant language effect on moral judgments for the footbridge dilemma, $\chi^2(1) = 4.25$, $p = .039$, $\varphi = .24$, 95% CI [.00, .47]. Importantly, however, this effect did not reflect the MFLE but the opposite pattern, i.e., a higher proportion of utilitarian choices in the native-language group (0.17) than in the foreign-language group (0.03). Finally, it is worth noting that our test on the aggregate data of all V1-dilemmas was sensitive enough to reliably detect a considerably smaller MFLE (n = 64, 1-β = .99, $w \geq .19$). Thus, it is rather unlikely that the null effect for the MFLE on the aggregate level is due to insufficient power, at least if assuming that the MFLE is not so small as to be essentially meaningless in real-world contexts.

3 Experiment 2

Experiment 2 aimed to address the limitations of Experiment 1 raised above. For this reason, we implemented the following main changes. First, we varied the different dilemma versions between subjects instead of within subjects. Second, each participant only had to judge four dilemmas. Finally, we collected a much larger sample to account for the different experimental design and to increase the sensitivity of our statistical tests. Experiment 2 was preregistered on March 20, 2019 with the template from AsPredicted.org on the Open Science Framework (https://osf.io/tdwvr).

3.1 Method

3.1.1 Participants

We aimed for a target sample size of 120 participants in each dilemma version condition, i.e., 480 in total. A power-analysis with G*Power (Faul et al., 2009) indicated that with this sample size, a χ^2-test (df = 1, α = .05) will have a power of .80 to detect a MFLE of size $w \geq .26$ for a single dilemma and $w \geq .13$ for an aggregated analysis of all four dilemmas in their standard form (V1).[3] Participants were recruited between March 24, and April 8, 2019 via social media and mailing lists for the web-based study. Because we expected a large number of data exclusions and did not want to miss the target sample size as in Experiment 1, we recruited as much participants as possible in the specified time frame. We did not apply a sample-based stopping rule and did not run any interim analyses.

A total of 673 participants completed the experiment. However, in line with our pre-registered exclusion criteria, we excluded 168 participants for the following reasons, some of which were concurrent: non-native German speaker (n = 22), accuracy below 50% in the comprehension test (n = 37), declaration of non-serious participation (n = 9), self-rated English proficiency as mother tongue or parent whose native language is English (n = 12), failure to pass the one-item attention check (n = 114). We had added the latter criterion because Experiment 2 was conducted as a web-based study and we only wanted to include participants who focused on the study. In contrast, we no longer excluded participants

3 By mistake, we had calculated the sensitivity of the χ^2-test based on 60 (instead of 120) participants per dilemma version in the pre-registration. The sensitivity analysis described above is the correct analysis.

who had spent more than ten months in an English-speaking country. We had dropped this criterion because—as already noted in the discussion of Experiment 1—it can be criticized for being arbitrary. The final sample comprised 505 native-German speakers (340 female, 156 male, and 9 diverse). More than half of the participants (65%) were students. The age of the participants ranged from 18 to 70 years (M = 27.6, SD = 9.7). Demographic information by condition is provided in the Appendix (see Table A1).

3.1.2 Materials

We selected four dilemmas from Experiment 1. Two dilemmas (immune deficiency, torture) were original CNI model dilemmas, the other two dilemmas (footbridge, terrorist) were dilemmas for which the MFLE had been reported at least once. We adapted some phrases of the selected dilemmas for the following reasons. First, we aimed at replacing words that are potentially unfamiliar to non-native English speakers with more familiar words. Second, we aimed at matching the dilemmas of both language versions as closely as possible. Because the four dilemma versions were manipulated between participants, all comprehension-test questions were adapted to the dilemma version of the respective group. All dilemmas and comprehension-test questions of Experiment 2 are available at: https://osf.io/4qg3c/

3.1.3 Design

The design was again a 2 (group: foreign language, native language) × 4 (dilemma version: V1, V2, V3, V4) design. Unlike in Experiment 1, both factors were manipulated between participants.

3.1.4 Procedure

The procedure was the same as in Experiment 1, except for the following changes: The experiment was implemented as a web-based study. In the moral-judgment phase, each participant saw only four dilemmas, each of which appeared in the same dilemma version, depending on the dilemma version group. The order of the four dilemmas was counterbalanced according to a Latin square. The questions of the comprehension test matched the respective dilemma version and appeared in the same order as the dilemmas in the moral-judgment phase. Subsequently, participants performed the LexTALE in the language version of their respective group. Unlike in Experiment 1, participants of both language groups were then asked several questions about their English skills (overall proficiency, ratings on the dimensions oral production, written production, oral comprehension, and written comprehension) and about their English learning background (whether one of their parents was a native English speaker, at which age they had started to learn English, how many months they had spent in an English-speaking country). Moreover, at the end of the experiment participants of both language groups were asked whether they had been familiar with any of the presented moral dilemmas.

3.2 Results

We had pre-registered the following analyses: 1) conventional dilemma analyses with χ^2-tests, 2) multinomial CNI model analyses, and 3) secondary analyses testing the robustness of our findings.

3.2.1 Conventional analyses

Similar to Experiment 1, we first compared the proportion of "utilitarian choices" between language groups. For better comparability with prior MFLE studies, this analysis was restricted to the V1-dilemma group (n = 124). As illustrated in Figure 3a, the proportion of utilitarian choices in response to the V1 dilemmas was slightly higher in the foreign-language group than in the native-language group. Thus, the pattern of results was in the predicted direction. But this

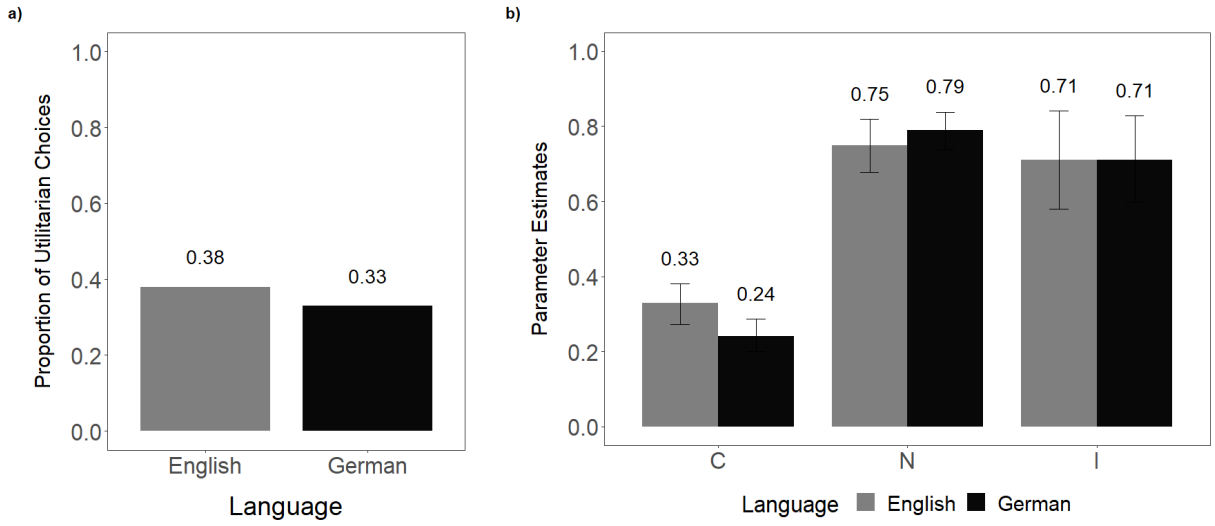

Figure 3: Results of a) the conventional analysis and b) the CNI analysis for the foreign-language group (English) and the native-language group (German) of Experiment 2.

observed group difference was not statistically significant, $\chi^2(1) = 1.46$, $p = .226$, $\varphi = .05$, 95% CI [.00, .14]. To test the dilemma-specificity of the MFLE, we repeated this analysis on the item level (see Table 4). We did not find a significant language effect for the immune deficiency, the footbridge, and the terrorist dilemma, χ^2s(1) ≤ 0.38, ps ≥ .536, φs ≤.06. However, there was a MFLE for the torture dilemma, $\chi^2(1) = 3.88$, $p = .049$, φs = .18, 95% CI [.00, .35], which is also the only dilemma for which we had found the effect in Experiment 1, at least descriptively.

3.2.2 CNI model analyses

As in Experiment 1, moral judgments for all dilemma versions were analyzed with the CNI model. When fitting the model to the aggregated data of all four dilemmas the goodness-of-fit test revealed a significant misfit between the predicted and observed response frequencies, $G^2(2) = 6.71$, $p = .035$. However, it must be noted that given the large sample size, the sensitivity of our goodness-of-fit test was very high. More precisely, a sensitivity analysis with G*Power revealed that given a significance level of $\alpha = .05$, the goodness-of-fit test had a power of .99 to detect deviations of the model as small as $w = .10$. Nevertheless, the model's parameter estimates (see Figure 3b) and the following statistical inferences have to be interpreted with caution given the poor model fit. Unlike in Experiment 1, participants' sensitivity to consequences was significantly higher in the foreign-language group compared to the native-language group, $\Delta G^2(1) = 5.52$, $p = .019$. In contrast, language neither affected participants' sensitivity to norms, $\Delta G^2(1) = 0.82$, $p = .367$, nor their general preference for inaction versus action, $\Delta G^2(1) < 0.01$, $p = .974$. As in Experiment 1, both language groups showed a preference for inaction as indicated by I-parameters significantly larger than .50, ΔG^2s(1) ≥ 9.04, $p \leq .003$.

We also fit the CNI model separately to each of the four different dilemmas. The model fit the data of the footbridge dilemma, the immune-deficiency dilemma, and the torture dilemma, G^2s(2) ≤ 4.42, $p \geq .110$, but not the data of the terrorist dilemma, $G^2(2) = 8.37$, $p = .015$. When comparing parameter estimates between language groups for each parameter of a dilemma, we did not observe any significant differences with one exception: The N-parameter for the immune deficiency dilemma was significantly higher in the native-language than in the foreign-language group, $\Delta G^2(1) = 4.59$, $p = .032$ (see Table A4 in the Appendix for parameter estimates per dilemma).

3.2.3 Secondary analyses

As pre-registered, we compared the data of the comprehension test between language groups to rule out that participants in the foreign-language group had difficulties to understand the presented dilemmas. As in Experiment 1, however, the

Table 4: Proportion of utilitarian choices for the standard version (V1) of each dilemma as a function of language (English = foreign language; German = native language) in Experiment 2.

Dilemma	English	German	χ^2-test (df = 1)	φ with .95 CI
Torture	.72	.54	$\chi^2 = 3.88$, $p = .049$	.18 [.00, .35]
Immune Deficiency	.22	.19	$\chi^2 = 0.11$, $p = .737$	.03 [.00, .20]
Footbridge	.24	.19	$\chi^2 = 0.38$, $p = .536$	.06 [.00, .23]
Terrorist	.35	.39	$\chi^2 = 0.17$, $p = .682$	.04 [.00, .21]

number of correct answers to the comprehension-test questions did not significantly differ between the native-language group ($M = 12.9$, $SD = 2.0$) and the foreign-language group ($M = 13.1$, $SD = 1.9$), $t(503) = 1.26$, $p = .209$, $d = 0.11$, 95% CI [-0.06, 0.29]. In addition, we again compared participants' LexTALE scores between groups. As expected, LexTALE scores were significantly higher in the native-language group (German LexTALE score: $M = 88.9$, $SD = 4.7$) than in the foreign-language group (English LexTALE score: $M = 80.7$, $SD = 10.8$), $t(262.39) = 10.31$, $p < .001$, $d = 0.99$, 95% CI [0.75, 1.12]. Hence, as in Experiment 1, participants' language proficiency was clearly higher in German than in English so that the absence of the MFLE cannot be attributed to equivalent proficiencies in both languages.

3.3 Discussion

As in Experiment 1, the MFLE did not replicate, at least not on the aggregate level. On the level of the individual dilemmas, however, the effect was present for the torture dilemma. Interestingly, this had also been the only of the eight dilemmas in Experiment 1 for which the descriptive pattern was in the expected direction. On the aggregate level, data analyses with the CNI model showed a significant larger *C*-parameter in the foreign-language compared to the native-language group. However, once again, we observed differences in the pattern of the parameter estimates when fitting the model per dilemma. The *C*-parameter was the only parameter that was consistently larger in the foreign-language group than in the native-language group across dilemmas, although this difference did not reach statistical significance when compared per dilemma. In contrast, language differences for the other two parameters were rather unsystematic and only one parameter, the *N*-parameter for the immune deficiency dilemma, differed significantly between language groups, with a higher estimate in the native-language compared to the foreign-language group. But note that this effect for the immune deficiency dilemma had not showed up in Experiment 1, not even descriptively.

4 General Discussion

In two pre-registered experiments, we investigated the robustness and the underlying processes of the MFLE. However, similar to several other studies, we did not find the effect of interest (e.g., Białek et al., 2019; Chan et al., 2016; Hayakawa et al., 2017; Mills & Nicoladis, 2020; Muda et al., 2017; Muda et al., 2020; Winskel & Bhatt, 2020). Likewise, model-based analyses with the CNI model did not reveal any systematic language effects on participants' sensitivity to consequences, their sensitivity to norms, or their inaction/action preference.

4.1 Implications for MFLE theories

Given the absence of the MFLE in our experiments and the inconsistent data patterns across the two experiments as well as across the different dilemmas, it is impossible to draw clear theoretical conclusions regarding the effect of foreign-language processing on moral judgments. For example, despite the higher *C*-parameter in the foreign-language group compared to the native-language group in Experiment 2, our findings do not provide unequivocal support for the

weighting hypothesis (i.e., the assumption that foreign-language processing increases the weighting of consequences). This is because foreign-language processing did not increase the *C*-parameter in Experiment 1. In this experiment, we even observed the opposite pattern, at least descriptively. When analyzing the data of Experiment 1 per dilemma, two dilemmas even displayed a significantly lower *C*-parameter in the foreign-language compared to the native-language group. Similarly, we neither obtained empirical support for the norm-accessibility hypothesis, which assumes a reduced accessibility of norms in a foreign language. If this was the case, we should have observed a lower sensitivity to norms in the foreign-language as compared to the native-language group. In contrast, however, neither Experiment 1 nor Experiment 2 displayed significant language differences in the *N*-parameter when fitting the CNI model to the aggregated data. Even though we did observe significant differences in the *N*-parameter in a few cases when analyzing the data per dilemma, the direction of these effects was inconsistent.

4.2 Compatibility with other MFLE studies

Comparisons with other model-based studies on the MFLE (Białek et al., 2019; Hayakawa et al., 2017; Hennig & Hütter, 2021; Muda et al., 2017) also show inconsistent findings, both when compared to our study and when compared among each other. The PDP studies by Muda et al. (2017) and Hayakawa et al. (2017) as well as the CNI-model study by Białek et al. (2019) found that foreign language reduces people's sensitivity to consequences (or their utilitarian inclination, respectively) as well as their sensitivity to norms (or their deontological inclination, respectively). However, a mini meta-analysis of Białek et al. also showed that the decreases of the PDP model's *U*-parameter (d = 0.22, 95% CI [0.06, 0.38]) and the CNI model's *C*-parameter (d = 0.27, 95% CI [0.11, 0.43]) were quite small. The same was true for the PDP model's *D* parameter (d = 0.20, 95% CI [0.04, 0.36]) and the CNI model's *N*-parameter (d = 0.29, 95% CI [0.14, 0.45]). In light of these effect-size estimates, our PDP analyses in Experiment 1 were clearly underpowered (d = 0.20, 1-β < .20). As for the CNI model analyses, we could not perform corresponding power analyses because these require estimates of the true parameters in the population. Yet, observed parameter values varied considerably in their magnitude and in the direction of the language effect. This was not only when comparing the results of Experiments 1 and 2, but also when comparing our results with the ones of the CNI-model study of Białek et al. For example, in Experiment 2, we found a significant increase of the *C*-parameter in the foreign-language group compared to the native-language group, which is in direct contrast to the effect observed by Białek and colleagues.

Discrepant results were also reported by Hennig and Hütter (2021) who investigated the MFLE with the proCNI model (Hennig & Hütter, 2020), an adapted version of the CNI model for dilemmas concerning proscriptive norms only.[4] The authors found no effect of foreign-language processing on participants' sensitivity to consequences, but a lower sensitivity to norms in a foreign-language as compared to a native-language group. Interestingly, however, this reduction in the *N*-parameter was only evident for high-involvement (i.e., personal) dilemmas, but not for low-involvement (i.e., impersonal) dilemmas. Note, however, that the interaction of involvement and language failed to reach statistical significance. What is more, the study by Hennig and Hütter (2021) was the first to indicate a language effect on behavioral response preferences captured by the proCNI model's *I* parameter. Yet, again, this effect was only found for the high-involvement dilemmas.

Taken together, empirical evidence regarding language effects on moral judgments is far from clear. Moreover, with a single exception, the MFLE did not replicate in the above-mentioned studies. Only Hennig and Hütter (2021) reported a significant language effect on participants' overall judgments. More precisely, in the first of their two experiments, participants in the foreign-language group were more inclined to judge norm-breaking behavior as acceptable compared to participants in the native-language group. However, because the authors aggregated judgments across incongruent *and* congruent dilemmas, it remains unclear if this finding truly reflects a MFLE (i.e., an increase of utilitarian choices in the foreign-language group for incongruent dilemmas in their standard form).

4 Because the proCNI model exclusively uses proscriptive dilemmas, the interpretation of its parameters is not fully equivalent to the CNI-model. Specifically, *N* refers to the endorsement of proscriptive norms and *I* reflects inertia, i.e., the tendency to stick to an already initiated action.

4.3 Potential moderators of the MFLE

The inconsistent findings on the MFLE and the underlying mechanisms of the effect allow for two different conclusions. Either there is no reliable MFLE, or the effect only occurs under very specific conditions. Considering the second possibility, we will discuss potential moderators of the effect below.

4.3.1 Personal/impersonal dilemmas

As already mentioned in the introduction section, the MFLE has been primarily reported for personal moral dilemmas (Costa et al., 2014; Geipel et al., 2015a), i.e., dilemmas in which the utilitarian option requires harming or killing one or more individuals by personal force. However, the personal/impersonal distinction cannot account for the full pattern of results. For example, Geipel et al. (2015a, Study 3) observed the MFLE for a low-emotion, impersonal dilemma (lost wallet dilemma) whereas the effect did not replicate for one of the high-emotion, personal dilemmas (crying baby dilemma). Similarly, we did not find the MFLE irrespective of whether the dilemmas involved personal force or not (see also Brouwer, 2019, Experiment 1; Chan et al., 2016; Muda et al., 2020). Moreover, we observed strong differences in the moral judgment patterns for the different dilemmas that cannot be fully attributed to the personal/impersonal distinction. It therefore seems worthwhile to examine further dilemma characteristics on moral judgments and the MFLE in future studies. To conclude, the relevance of the personal/impersonal distinction for the MFLE is not yet fully clear. Possibly, this variable moderates the MFLE under certain conditions only. For example, as emotional responsiveness to the personal dilemmas might approach that of native speakers with increasing language proficiency, it is conceivable that the MFLE is more likely to occur when foreign-language proficiency is low.

4.3.2 Language proficiency

In fact, a study by Čavar and Tytus (2018) suggests that a high level of foreign-language proficiency and a high degree of acculturation in the second language prevent the MFLE. However, because the study did not involve a control group, its finding allows for alternative interpretations such as a generally low replicability of the MFLE irrespective of language proficiency. In our study, for example, we did not find a MFLE even though our participants were clearly less proficient in the foreign language compared to the native language as indicated by an objective measure of language proficiency. Yet there are other studies whose findings suggest a link between language proficiency and the MFLE. For example, in Costa et al.'s (2014) study, participants who reported lower levels of foreign-language proficiency showed a larger MFLE for the footbridge dilemma than participants who reported higher levels of foreign-language proficiency. Similarly, Corey et al. (2017) observed a significant negative correlation between language proficiency and utilitarian judgments for the lost wallet dilemma. Interestingly, Brouwer (2019) only replicated the MFLE when the dilemmas were presented auditorily, but not when they were presented in written format (but see Muda et al., 2020). Possibly, this finding relates to the fact that oral comprehension is typically worse than written comprehension. Yet, in a recent meta-analysis by Circi et al. (2021), language proficiency did not moderate the MFLE, which was small overall, $g = 0.22$, 95% CI [0.14, 0.30]. It should be noted, however, that the 38 studies on which the meta-analysis was based, used different language-proficiency measures, most of which were based on self-report instead of standardized, objective measures.

4.3.3 Linguistic similarity

In a recent study by Dylman and Champoux-Larsson (2020), the MFLE did not replicate for Swedish participants when reading the dilemmas in English or Norwegian, but when reading the dilemmas in French. The authors argued that the MFLE does not replicate if native and foreign language share a high linguistic similarity or if the foreign language is learned in an emotion-rich context (e.g., TV, shows, music). However, given the fact that the self-reported reading comprehension was lower in French compared to the other languages, it might nevertheless be language proficiency rather than linguistic similarity that drove the effect. Yet, the meta-analysis by Circi et al. (2021) also supports the

linguistic-similarity hypothesis. More specifically, for studies that had examined the effect with linguistically dissimilar languages (e.g., English and Spanish), the estimated effect-size of the MFLE was $g = 0.30$, 95% CI [0.21, 0.39]. In contrast, for linguistically similar languages (e.g., English and German), the effect size was only $g = 0.06$, 95% CI [-0.05, 0.18] and no longer different from zero. Although this language-similarity hypothesis can account for the absence of the MFLE in our experiments, the hypothesis has a serious shortcoming: the lack of theoretical foundation. That is, there is no theoretical rational why language similarity should affect moral decision making whereas language proficiency does not. Besides, related factors such as the age of foreign-language acquisition, the frequency of language use, and the context of language use (private vs. academic context) have typically not been controlled for in studies on the MFLE. What is more, Circi and colleagues may have overlooked another variable that was possibly confounded with language similarity in their meta-analysis: the number of presented dilemmas.

4.3.4 Number of presented dilemmas

One striking observation is that almost all studies that used a model-based approach to investigate language effects on moral judgments failed to replicate the MFLE (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2017). As in our first experiment, participants in these studies had to read and judge a large number of dilemmas because the modeling required judgments of different versions of each dilemma. Other studies on the MFLE, in contrast, typically relied on just a few dilemmas. In contrast, however, a study by Chan et al. (2016) involved a battery of 39 dilemmas in total. Although the tested languages in this study were linguistically dissimilar (native language: Chinese; foreign language: English), the MFLE did not replicate (except for the footbridge dilemma). At present, the idea that an increasing number of dilemmas might counteract the MFLE is speculative in nature and requires further investigation. Yet, a comparison of our findings from Experiment 1 (based on 32 dilemmas per person) and Experiment 2 (based on 4 dilemmas per person) also points in this direction, albeit only descriptively. Moreover, on a theoretical level, it makes sense to assume that individuals respond less emotionally when judging several moral dilemmas in a row so that the dilemma processing of native and nonnative speakers might converge after only a few trials. Future studies should investigate this point more systematically to figure out whether order effects truly matter in the context of the MFLE.

5 Conclusion

In two experiments we investigated the MFLE with a model-based approach but did not find the effect of interest in the first place. Instead, we found very inconsistent language effects on moral judgments both, within and across our two experiments. It is possible that these different patterns of results are related to the different dilemma sets used in each study and/or relate to others contextual factors discussed above. We therefore conclude that if there is a MFLE, the effect is quite small and/or very fragile and context specific. Consequently, the significance of the effect in the real world is highly questionable.

Data Availability Statement: The datasets and materials of Experiments 1 and Experiment 2 are available on the Open Science Framework at: https://osf.io/b2as9/.

Ethics Statement: This study was carried out in accordance with the ethical guidelines of the German Psychological Society (DGPs). All subjects gave their informed consent in accordance with the Declaration of Helsinki.

Conflict of Interests: The authors have no conflicts of interest to disclose.

Acknowledgements: This research was supported by the “Fair@UMA program” of the University of Mannheim and the “Brigitte-Schlieben-Lange program” of the Ministry of Science, Research, and the Arts Baden-Württemberg.

References

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. https://doi.org/10.1016/j.jesp.2019.103855

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. https://doi.org/10.1080/01434632.2019.1585863

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. https://doi.org/10.1080/01434632.2017.1304397

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*(1), 55–65. https://doi.org/10.1111/ajsp.12123

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

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

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216–235. https://doi.org/10.1037/a0031021

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), 1109–1128. https://doi.org/10.1037/xlm0000356

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Haefner, J., & Keysar, B. (2014). Your morals depend on language. *PloS One*, *9*, e94842. https://doi.org/10.1371/journal.pone.0094842

Crone, D. L., & Laham, S. M. (2017). Utilitarian preferences or action preferences? De-confounding action and moral code in sacrificial dilemmas. *Personality and Individual Differences*, *104*, 476–481. https://doi.org/10.1016/j.paid.2016.09.022

Dylman, A. S., & Champoux-Larsson, M.-F. (2020). It's (not) all Greek to me: Boundaries of the foreign language effect. *Cognition*, *196*, 104148. https://doi.org/10.1016/j.cognition.2019.104148

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343-376. https://doi.org/10.1037/pspa0000086

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2016). Understanding responses to moral dilemmas: Deontological inclinations, utilitarian inclinations, and general action tendencies. In J. P. Forgas, L. Jussim, & A. M. van Lange (Eds.), *Social psychology of morality* (pp. 91–110). Psychology Press.

Gawronski, B., Conway, P., Hütter, M., Luke, D. M., Armstrong, J., & Friesdorf, R. (2020). On the validity of the CNI model of moral decision-making: Reply to Baron and Goodwin (2020). *Judgment and Decision Making*, *15*(6), 1054–1072. http://journal.sjdm.org/20/200813/jdm200813.pdf

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. https://doi.org/10.1371/journal.pone.0131529

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

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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Harris, C. L. (2004). Bilingual speakers in the lab: Psychophysiological measures of emotional reactivity. *Journal of Multilingual and Multicultural Development*, *25*(2-3), 223–247. https://doi.org/10.1080/01434630408666530

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. https://doi.org/10.1177/0956797617720944

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology*, *118*(1), 22–56. https://doi.org/10.1037/pspa0000173

Hennig, M., & Hütter, M. (2021). Consequences, norms, or willingness to interfere: A proCNI model analysis of the foreign language effect in moral dilemma judgment. *Journal of Experimental Social Psychology*, *95*, 104148. https://doi.org/10.1016/j.jesp.2021.104148

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*(1), 7–31. https://doi.org/10.1007/s11336-004-1188-3

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

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*. Advance online publication. https://doi.org/10.1080/01434632.2020.1813145

Moore, A. B., Lee, N. Y. L., Clark, B. A. M., & Conway, A. R. A. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making*, *6*, 186–195.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Muda, R., Niszczota, P., Bialek, M., & Conway, P. (2017). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(2), 321–326. https://doi.org/10.1037/xlm0000447

Muda, R., Pieńkosz, D., Francis, K. B., & Białek, M. (2020). The moral foreign language effect is stable across presentation modalities. *Quarterly Journal of Experimental Psychology*, *73*(11), 1930–1938. https://doi.org/10.1177/1747021820935072

Singmann, H., & Kellen, D. (2013). Mptinr: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. https://doi.org/10.3758/s13428-012-0259-0

Thoma, D., & Baum, A. (2019). Reduced language processing automaticity induces weaker emotions in bilinguals regardless of learning context. *Emotion*, *19*(6), 1023–1034. https://doi.org/10.1037/emo0000502

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://www.jstor.org/stable/27902416

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995

Winskel, H., & Bhatt, D. (2020). The role of culture and language in moral decision-making. *Culture and Brain*, *8*, 207–225. https://doi.org/10.1007/s40167-019-00085-y

Appendix

Table A1: Demographic information by language group (English = foreign language; German = native language) in Experiments 1 and 2.

	Experiment 1		Experiment 2	
	English	**German**	**English**	**German**
Final sample size	*n* = 28	*n* = 36	*n* = 208	*n* = 297
% Female	75%	83%	65%	69%
Mean (*SD*) age	22.1 (3.2)	21.8 (2.6)	27.5 (9.6)	27.7 (9.7)

Note. The reported data refer to the final sample sizes, i.e., after application of exclusion criteria.

Table A2: Reported English proficiency in the foreign-language groups in Experiments 1 and 2.

	Experiment 1	Experiment 2
Self-assessed English proficiency level (*n*):		
Beginner	0	1 (<1%)
Basic	2 (7%)	8 (4%)
Conversational	4 (14%)	42 (20%)
Fluent	20 (71%)	61 (29%)
Close to Native	2 (7%)	29 (14%)
Missings	0	67 (32%)
Mean (*SD*) of self-rated English skills:		
Oral comprehension	5.46 (0.84)	4.85 (1.13)
Oral production	4.86 (1.08)	4.87 (1.08)
Written comprehension	5.25 (1.14)	5.72 (0.98)
Written production	4.75 (0.97)	5.57 (1.03)
English certificate, e.g., TOEFL (*n*):		
Yes	16 (57%)	----
No	12 (43%)	---
Mean (*SD*) age of acquisition	8.2 (2.0)	9.3 (2.9)
Mean (*SD*) time spent in English-speaking country (in months).	2.4 (2.7)	4.8 (8.5)

Note. The reported data refer to the final sample sizes, i.e., after application of exclusion criteria.

Table A3: Parameter estimates of the CNI model with 95% CIs for each dilemma as a function of language (English = foreign language; German = native language) in Experiment 1.

Dilemma	Parameter	English	German	χ^2-test (df = 1)
Abduction	*C*	.32 [.14, .51]	.65 [.54, .76]	$\chi^2 = 9.75$, $p = .002$
	N	.21 [.00, .49]	.66 [.37, .94]	$\chi^2 = 4.66$, $p = .031$
	I	.50 [.32, .67]	.62 [.00, 1.0]	$\chi^2 = 0.29$, $p = .591$
Transplant	*C*	.18 [.04, .32]	.06 [.00, .19]	$\chi^2 = 1.34$, $p = .247$
	N	.74 [.56, .91]	.44 [.28, .61]	$\chi^2 = 5.23$, $p = .022$
	I	.55 [.00, 1.0]	.58 [.43, .73]	$\chi^2 = 0.02$, $p = .880$
Torture	*C*	.70 [.56, .81]	.57 [.44, .69]	$\chi^2 = 2.12$, $p = .146$
	N	.89 [.66, 1.0]	.79 [.59, 1.0]	$\chi^2 = 0.41$, $p = .522$
	I	.00 [.00, 1.0]	.73 [.00, 1.0]	$\chi^2 = 2.05$, $p = .152$

Continued **Table A3:** Parameter estimates of the CNI model with 95% CIs for each dilemma as a function of language (English = foreign language; German = native language) in Experiment 1.

Dilemma	Parameter	English	German	χ^2-test (df = 1)
Assisted Suicide	*C*	.61 [.46, .75]	.67 [.54, .78]	$\chi^2 = 0.36$, $p = .549$
	N	.15 [.00, .54]	.24 [.00, .61]	$\chi^2 = 0.14$, $p = .709$
	I	.73 [.51, 1.0]	.47 [.21, .79]	$\chi^2 = 2.36$, $p = .124$
Immune Deficiency	*C*	.14 [.00, .29]	.18 [.04, .32]	$\chi^2 = 0.16$, $p = .693$
	N	.76 [.61, .95]	.66 [.50, .83]	$\chi^2 = 0.76$, $p = .383$
	I	.73 [.00, 1.0]	.62 [.37, .91]	$\chi^2 = 0.37$, $p = .541$
Vaccine	*C*	.32 [.15, .48]	.39 [.25, .53]	$\chi^2 = 0.41$, $p = .521$
	N	.52 [.28, .75]	.32 [.08, .57]	$\chi^2 = 1.28$, $p = .259$
	I	.65 [.42, 1.0]	.55 [.37, .75]	$\chi^2 = 0.46$, $p = .496$
Footbridge	*C*	.03 [.00, .11]	.16 [.06, .28]	$\chi^2 = 4.20$, $p = .040$
	N	.90 [.80, .99]	.89 [.77, 1.0]	$\chi^2 = 0.02$, $p = .900$
	I	1.0 [.87, 1.0]	.72 [.00, 1.0]	$\chi^2 = 1.26$, $p = .261$
Terrorist	*C*	.36 [.21, .51]	.53 [.41, .66]	$\chi^2 = 2.91$, $p = .088$
	N	.80 [.62, 1.0]	.77 [.57, 1.0]	$\chi^2 = 0.05$, $p = .829$
	I	.78 [.00, 1.0]	1.0 [1.0, 1.0]	$\chi^2 = 1.62$, $p = .203$

Note: CIs were estimated using non-parametric bootstrapping with $n = 1000$ bootstrap samples.

Table A4: Parameter estimates of the CNI model with 95% CIs for each dilemma as a function of language (English = foreign language; German = native language) in Experiment 2.

Dilemma	Parameter	English	German	χ^2-test (df = 1)
Torture	*C*	.57 [.46, .67]	.46 [.37, .54]	$\chi^2 = 2.50$, $p = .114$
	N	.75 [.54, .94]	.82 [.69, .93]	$\chi^2 = 0.45$, $p = .503$
	I	.29 [.00, 1.0]	.68 [.33, 1.0]	$\chi^2 = 2.40$, $p = .121$
Immune Deficiency	*C*	.23 [.11, .33]	.15 [.07, .23]	$\chi^2 = 1.07$, $p = .301$
	N	.56 [.44, .69]	.74 [.64, .84]	$\chi^2 = 4.59$, $p = .032$
	I	.90 [.78, 1.0]	.83 [.67, 1.0]	$\chi^2 = 0.63$, $p = .428$
Footbridge	*C*	.13 [.03, .22]	.08 [.00, .16]	$\chi^2 = 0.60$, $p = .440$
	N	.78 [.66, .89]	.78 [.69, .88]	$\chi^2 < 0.01$, $p = .995$
	I	.62 [.36, 1.0]	.56 [.36, .80]	$\chi^2 = 0.17$, $p = .680$
Terrorist	*C*	.36 [.26, .46]	.28 [.20, .37]	$\chi^2 = 1.37$, $p = .243$
	N	.87 [.76, .98]	.83 [.73, .91]	$\chi^2 = 0.35$, $p = .552$
	I	.60 [.00, 1.0]	.84 [.59, 1.0]	$\chi^2 = 0.96$, $p = .328$

Note: CIs were estimated using non-parametric bootstrapping with $n = 1000$ bootstrap samples.