Cognition xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Thinking in a foreign language distorts allocation of cognitive effort: Evidence from reasoning

Michał Białek[a,c,*], Rafał Muda[b], Kaiden Stewart[a], Paweł Niszczota[d], Damian Pieńkosz[b]

[a] Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada
[b] Faculty of Economics, Maria Curie-Skłodowska University, Plac Marii Skłodowskiej-Curie 5, Lublin, Poland
[c] Institute of Psychology, University of Wroclaw, Dawida 1, Wroclaw, Poland
[d] Department of International Finance, Poznań University of Economics and Business, al. Niepodległości 10, Poznań, Poland

ARTICLE INFO

Keywords:
Foreign Language Effect
Reasoning
Logical intuition
Dual-process theory
Conflict detection
Signal detection theory

ABSTRACT

Bilinguals, in their foreign language, are spared from several decision-making biases. We examined this "Foreign Language Effect" in the context of logical reasoning, in which reasoners are required to track the logical status of a syllogism, ignoring its believability. Across three experiments, we found the reverse Foreign Language Effect; foreign language reasoners are less able to evaluate the logical structure of syllogisms, but no less biased by their believability. One path to succeeding in reasoning tasks is always engaging in reflective processing. A more efficient strategy is metacognitively tracking whether belief-based intuitions conflict with logic-based intuitions and only reflecting when such conflict is present. We provide evidence that foreign language reasoners are less accurate because they struggle to detect belief-logic conflict, and in turn fail to engage in reflective processing when necessary to override the incorrect, intuitive response. We propose that foreign language reasoners are less able to detect belief-logic conflict either due to weakened intuitions or due to a more conservative threshold for the detection of conflict between multiple competing intuitions. Data for the experiments can be accessed publicly at https://osf.io/phbuq/

1. Introduction

Research on judgement and decision making suggests that bilinguals, when making decisions in their foreign language (FL), are protected from several of the biases they would otherwise be affected by in their native language (NL). For example, FL decision-makers tend to be less affected by the framing of the decision (Keysar, Hayakawa, & An, 2012), loss aversion (Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014), hot hand fallacy (Gao, Zika, Rogers, & Thierry, 2015), and superstitious beliefs (Hadjichristidis, Geipel, & Surian, 2019). This change in susceptibility to decision-making biases is dubbed the Foreign Language Effect (FLE).

To this point, two main explanations have been proposed as to what mechanism is responsible for the FLE. The first of these explanations proposes that using a foreign language attenuates the emotional resonance elicited by a problem, as the foreign language is usually acquired in a low-emotion, classroom context. Therefore, FL decision-makers might not engage emotional processing as easily as NL decision-makers and in turn be less susceptible to emotionally-based biases. The second explanation proposes that using a foreign language promotes deliberation, as processing in a foreign language is less fluent. In this case, processing difficulties serve as a signal that more deliberative processes are needed and thus, the role of intuition is reduced (Costa, Vives, & Corey, 2017; Hayakawa, Costa, Foucart, & Keysar, 2016). Obviously, these mechanisms are not mutually exclusive, and the FLE could be a product of both mechanisms.

In a direct test of the emotional explanation, the FLE in moral judgments was not mediated by emotionality (Geipel, Hadjichristidis, & Surian, 2015), and emotionality was similar across groups that demonstrated multiple types of common decision-making fallacies (Miozzo et al., 2020). The alternative, cognitive explanation also fails to fully account for the effect: no language effects were observed in the cognitive reflection tests (Białek, Paruzel-Czachura, & Gawronski, 2019; Costa, Foucart, Amon, et al., 2014; Mækelæ & Pfuhl, 2019), nor in the deliberative use of statistical information in gambling decisions (Muda, Walker, Pieńkosz, Fugelsang, & Białek, 2020), or in intertemporal choice (Białek, Domurat, Paruzel-Czachura, & Muda, 2020). There is some hint that cognitive reflection, a trait of how individuals allocate their cognitive effort, interacts with the FLE. Specifically, the effects of cognitive reflection are only visible when deciding in one's

* Corresponding author at: University of Wroclaw, Dawida 1, Wroclaw, Poland.
E-mail address: michal.bialek3@uwr.edu.pl (M. Białek).

https://doi.org/10.1016/j.cognition.2020.104420
Received 12 July 2019; Received in revised form 23 July 2020; Accepted 27 July 2020

native but not in one's foreign language (Białek et al., 2019, 2020).

Finally, recent research has complicated the understanding of the FLE, finding no effects in linguistically similar languages (Dylman & Champoux-Larsson, 2020; Mækelæ & Pfuhl, 2019), but observing a FLE in Italian local dialects (Miozzo et al., 2020). A report shows no effects of language in highly acculturated individuals (Cavar & Tytus, 2018, but see Białek & Fugelsang, 2019 for a commentary on why the evidence for this claim is weak). Hence, some linguistic and cultural factors certainly play a role in driving the FLE. Considering all the above, we lack a good explanation for the FLE and its cognitive mechanisms.

In the present work, we use a Dual-Process Theory framework to test the relative claims made by each of these mechanisms. In a Dual-Process Theory framework, it is assumed that cognition consists of two complementary processes: Type I processes are autonomous and do not require access to working memory, and Type II processes require access to working memory to cognitively decouple[1] and run mental simulations (Evans & Stanovich, 2013). The most prominent model of Dual-Process Theory proposes that, by default, people tend to use only Type I processes and engage in Type II reasoning when necessary (e.g., when they detect a conflict between two or more intuitive responses,[2] or if no intuitive response is readily available; Bago & De Neys, 2017; De Neys, 2014; De Neys & Pennycook, 2019; Handley & Trippas, 2015; Pennycook, Fugelsang, & Koehler, 2015; Šrol & De Neys, 2020; Stanovich, 2018; Thompson, Turner, & Pennycook, 2011; Trippas, Thompson, & Handley, 2017; Trippas & Handley, 2017). When reflecting, reasoners either try to investigate the logical status of the premises so that they can arrive at a valid conclusion, or they try to rationalize the most promising intuition (Pennycook et al., 2015). From this framework, one could postulate several possible mechanisms responsible for better performance on decision-making tasks. Perhaps individuals perform better when thinking in their FL because they, simply by encountering the problem in their FL, are forced to engage Type II processing. Alternatively, foreign-language reasoning could prompt conflict, either because reasoning in a foreign language produces the experience of conflict by default, or because reasoners produce no convincing intuitions, and the only route left is to solve the problem reflectively.

In this work, we investigate the foreign-language debiasing effect and its potential mechanisms. We focused on reasoning because the Dual-Process Theory originated and is most strongly understood in the domain of reasoning (De Neys, 2006; De Neys & Białek, 2017; Evans & Stanovich, 2013; Thompson & Johnson, 2014). In a typical reasoning task, participants are presented with logical syllogisms (e.g., “All A are B. All B are C. Therefore, all A are C”).[3] Participants are asked to evaluate the conclusion of the syllogism, indicating whether the conclusion does (or does not) follow logically and necessarily from the premises, while ignoring whether the participant believes the premises and conclusions are true in the real world. This paradigm allows researchers to examine the extent to which participants base their responses on the logical structure of the syllogism (as syllogisms differ in terms of logical validity) and the extent to which this process is affected by feelings of belief or disbelief (as syllogisms can differ in terms of the believability of their conclusions[4]). In order to detect the effects of believability, syllogisms are counterbalanced in terms of their objective properties such that sometimes belief and logic cue the same response (congruent trials) and sometimes they cue opposite responses (incongruent trials). For example, a syllogism can be valid and believable (congruent trial, where responding along either dimension prompts a ‘yes’ response) or valid and unbelievable (incongruent trial, where responding along the logical dimension cues a ‘yes’ response, but responding along the belief dimension cues a ‘no’ response). Whether and when participants, when solving syllogisms, correctly classify them as congruent or incongruent is an empirical question we pursue in this investigation.

Evaluating conclusions on the basis of belief is assumed to be intuitive and a product of Type I processing, while evaluation of the logical structure of the syllogism, although perhaps accomplished via logical intuition, is usually assumed to require Type II engagement (Evans & Stanovich, 2013; Pennycook et al., 2015).[5] Since one always first engages Type I processing and logic-based intuitions are usually weaker than belief-based intuitions, individuals typically consider the belief-based intuition first, and responding in line with logic then requires one to override the belief-based response with Type II processing. In other words, a response in line with believability suggests a failure to override a Type I response with a Type II response (Handley & Trippas, 2015).

The most dominant view in the field is that Type II responses are triggered by the detection of a conflict between multiple intuitions. More specifically, when assessing a syllogism, people automatically produce a Type I intuition related to the believability of the syllogism and a competing Type I intuition about the logical status of the syllogism. The belief-based, but not the logic-based intuitions are affective (Klauer & Singmann, 2013; Morsanyi & Handley, 2012). The belief-based intuition simply reflects one's assessment of the believability of the conclusion and premises. The logic-based intuition reflects one's assessment of the logical validity of a conclusion, and can reflect innate or learned logical rules, or some logical heuristic (e.g., the atmosphere heuristic: promoting affirmative conclusions if the premises are affirmative or promoting particular conclusions if any of the premises is particular rather than universal; Woodworth & Sells, 1935). This logical intuition has to be distinguished from effortful processing of Type II, which fleshes out the logical structure of the premises and only then computes the validity of the conclusion. Reasoners have access to the output of both Type I and Type II processes, but have access to the content of only the Type II processing. In other words, for intuitive (i.e., Type I) judgments, reasoners will have an intuitive sense of whether a conclusion is valid or invalid, but for reflective (i.e., Type II) judgments, they also know *why* they believe so. Type I logical intuitions and Type II

[1] Cognitive decoupling is defined as the ability to distinguish supposition from belief and to aid rational choices by running thought experiments (Evans & Stanovich, 2013; Stanovich, 2009).

[2] Dual-process theories of reasoning have suffered circularity problems. They sometimes require that Type II processes will be engaged if some other process detects the Type I output is incorrect. This is problematic if such a monitoring process is assumed to be Type II in and of itself. That is, it is strange to postulate that Type II processes effectively trigger themselves. By conceptualizing the process as including belief-based and logic-based intuitions, the newer models of reasoning allow instead for the possibility that lower level (i.e., not Type II) processes detect conflict between belief-based and logic-based intuitions. Type II processing in such models is required to resolve the conflict, or to double-check the intuitive output.

[3] The middle-term “B” is redundant to assess both the believability and the validity of a conclusion. For believability, this is because a participant is informed that both sentences including it (premises) are true, and the conclusion does not include the middle-term. For validity judgments the B term is also redundant, because the task is to assess the logical status of a conclusion which consists of terms A and C. Hence, the “B” term is sometimes substituted with an abstract word because it helps to control the believability of premises. We adopted this strategy in the current project.

[4] When reasoning about in the real world, premises and conclusions refer to real-world objects, and can be intuitively judged to be believable or unbelievable. The believability of a syllogism can be therefore a product of the believability of the premises and the conclusion. To avoid this complexity, some experiments use syllogisms that use a-true premises like “all dogs are q; some q bark”. These premises have no reference categories in the real world, and only produce logical, but not belief-based intuitions. In such cases, the conclusion is the only source of believability.

[5] In cases where the premises and the conclusion are a-true, one can arrive to a correct response via Type I processes because no belief-based intuition is produced (Bago & De Neys, 2017). To simplify, we further describe reasoning as if all conclusions would refer to categories which can be believable or not.

logical processing are not identical in every way. However, problems typically used in reasoning research have been simplified so that both logics (Type I and Type II) must agree. As such, belief-logic congruence necessarily refers to the agreement between belief and both types (Type I and Type II) of logic.

If either the belief-based or logical intuition is much stronger than the other, no conflict will be detected and the output will simply be the dominant intuitive response (usually the belief-based one; De Neys, 2006; De Neys, 2014). If, however, these intuitions are of approximately equal strength, conflict will be detected, triggering Type II processing. In such cases, individuals will either attempt to override a Type I output with a Type II output or simply rationalize the most promising intuition (Handley & Trippas, 2015; Pennycook et al., 2015). To respond to each syllogism correctly, one strategy is to focus only on the structure of the syllogism, ignoring the believability of the conclusion. A more efficient system would intuitively estimate whether believability will cue a correct response (which is the case for congruent trials), or would not (which is the case for incongruent trials), and only engage in reflection for the latter type of syllogism. This minimizes the number of trials on which one must engage reflective, Type II processing without a proportional decline in accuracy (Stanovich, 2018).

One consequence of strong intuitions about believability is that logical validity judgments are biased by the believability of the syllogism (mostly cued by the believability of its conclusion, but also by its premises, see Solcz, 2011). This effect is labelled belief bias. This appears as a response bias: an overall tendency to endorse believable as opposed to unbelievable syllogisms as valid, regardless of their logical status (Dube, Rotello, & Heit, 2010). Because believability is usually assessed intuitively, belief bias is currently best thought of as an effect on Type I processing and not on Type II processing (Dube et al., 2010; Trippas et al., 2018; Trippas, Handley, & Verde, 2013).

The potential benefit of reasoning in one's foreign language is for two, non-exclusive reasons: people reflect more or are less biased by the believability of a conclusion. If FL reasoners experience an increase in reflective processing but not a decrease in intuitive processing, we should see in a syllogism task that they are more sensitive to the logical structure of the syllogism, but no more biased by its believability. If FL reasoners experience weaker affective resonance to the believability but are no more reflective, we should see in a syllogism task that they are no more sensitive to the logical structure of the syllogism, but less biased by its believability.

1.1. Overview of experiments

In three similarly designed experiments, we presented participants with 32 syllogisms, either in their native language or in their foreign language. Half of these syllogisms were valid, and half invalid. Half of each type had believable conclusions, and the other half unbelievable conclusions. As such, half of the syllogisms were congruent (i.e., logic and belief cue the same response) and half were incongruent (i.e., logic and belief cue different responses).

In all three experiments, we observed a decline in reasoning accuracy in one's FL. While still observing a FLE, ours is opposite to the direction typically reported. This decline in accuracy was driven by a decrease in the sensitivity to the logical structure of the syllogism, but not by greater reliance on the believability of the conclusion. This decrease in reasoning accuracy is caused by lower sensitivity to logical structure driven either by a lack of deliberation, or by distorted deliberation. Further analysis of our data revealed that NL reasoners were slower and less confident when they answered incongruent syllogisms incorrectly compared to answering congruent syllogisms correctly. Both of these responses are consistent with believability of a conclusion. This decrease in confidence suggests the detection of a logic-belief conflict, so that participants, even when not following logic, were still intuitively affected by it when responding. This conflict detection allows them to engage in reflection only when such conflict is strong enough, and results in reasonably high accuracy of reasoning. FL reasoners showed fewer signs of such conflict detection and accordingly failed to engage in deliberation when required. Altogether, we conclude that thinking in a foreign language disrupts conflict detection so that FL reasoners are less aware of when reflection is required. In turn, FL participants allocate their cognitive effort less accurately.

2. Experiment 1

2.1. Participants

Ultimately, we analyzed data from 129 participants (n = 104 female, M_{AGE} = 20, SD = 0.96). We originally recruited 209 participants from UMCS University in Lublin in exchange for $5 compensation. We dropped the data from participants who reported understanding of the materials to be lower than 5 on a 10-point scale (n = 36 in NL condition, and n = 26 in FL condition) (see e.g. Costa, Foucart, Hayakawa, et al., 2014; Muda, Niszczota, Białek, & Conway, 2018 for similar data reduction policy in FLE research), and a further n = 18 whose reasoning accuracy was below chance (see e.g. Handley, Newstead, & Trippas, 2011; Trippas, Handley, & Verde, 2014 for similar data reduction policy in reasoning research). The understanding criterion prevented participants with poor proficiency from contaminating our data. The accuracy criterion excluded inattentive participants. There was substantial overlap between these criteria; among those 62 participants who reported low understanding of the problems, the accuracy of 26% of them was below chance. For comparison, among those 147 who reported acceptable understanding, accuracy of only 12% of them was below chance. Hence, scoring below chance is a strong indicator of low comprehension of the task. As reported in an analysis of the full dataset, our findings remain consistent if these participants were not removed.

2.2. Materials and procedure

In this and all subsequent experiments participants were tested individually in lab, on PC's. This, and all other surveys in this research were created in LimeSurvey (Schmitz, 2010). Participants were assigned to one of two experimental conditions: the NL condition wherein they read and answered the reasoning problems in their first language (i.e., Polish), and the FL condition wherein they read and answered the problems in their second language (i.e., English). In each condition, participants judged the logical validity of 32 relatively non-complex syllogisms adopted from Trippas, Verde, and Handley (2014), in which validity and believability of the conclusion was manipulated. Conclusions in 16 of the syllogisms were valid and conclusions in the other 16 were invalid; half of each type of syllogism had believable conclusions and the other half had unbelievable conclusions. The task was to assess validity of a conclusion ignoring its believability using a yes-no response followed by a 3-point confidence rating (Not at all, moderately, very). We recorded response times for all validity assessments.

To control for believability, we assured only the conclusion could be evaluated along this dimension. We achieved this by introducing an abstract term to the premises, e.g.:

Some foxtrots are *jundors*
All *jundors* are dances
Therefore, some dances are foxtrots

Note that the middle term (*italicized* word in the example above connecting the premises, but not included in the conclusion) is a made-up pseudo-word. We highlighted this term in red and instructed participants to assume this word has no meaning. We adopted such procedure from previous research (Handley et al., 2011; Trippas et al., 2013; Trippas, Verde et al., 2014). This is critical for controlling the source of believability judgments. With middle terms being pseudo-words, the

premises are a-believable, i.e., no believability judgments can be made about them. However, each conclusion can be judged as believable or as unbelievable. Hence, the believability judgments can be only based on the believability of the conclusion. Without adequate control, the believability of premises interferes with the believability of a conclusion even if the participants are informed to assume the premises are true (Crane, 2016; Solcz, 2011).

2.3. Results and discussion

2.3.1. Traditional analysis

For all but ROC analyses, we used a free software JASP (JASP Team, 2020) and plotted the results with R package ggplot2 (Wickham, 2011).

To assess accuracy, we combined dichotomous yes-no responses with provided confidence ratings so that a correct response with high confidence was recoded into "6", with moderate confidence into "5", and with low confidence into "4". Similarly, an incorrect response with low confidence was recoded into "3", with moderate confidence into "2", and with high confidence into "1". Reaction times were log-transformed to reduce skew.

Having done this, we compared the accuracy of participants across experimental conditions (NL vs. FL). Contrary to what could have been expected from past research, we found that accuracy in the NL condition was higher than in the FL condition, $t(127) = 4.76$, $p < .001$, $d = 0.86$, 95% CI [0.49, 1.23] (Fig. 1). Participants spent the same time on solving the problems regardless of the experimental condition they were assigned to, $t(127) = 0.58$, $p = .563$, $d = 0.10$, 95% CI [−0.24, 0.45].

Past research on the FLE suggests that FL decision-makers benefit over (Costa, Foucart, Hayakawa, et al., 2014; Keysar et al., 2012), or at worst, show no difference from NL decision-makers (Hayakawa, Lau, Holtzmann, Costa, & Keysar, 2019; Mækelæ & Pfuhl, 2019; Muda, Walker, et al., 2020; Vives, Aparici, & Costa, 2018). We extended these claims into the reasoning domain. To our knowledge, the present study is the first to show a detriment in decision-making for individuals using their FL. In reasoning, the FLE appears in the opposite direction. That is, reasoning in one's FL is a disadvantage.

One problem with drawing inferences from traditional analyses is that we cannot fully dissociate sensitivity to logical structure from bias on the basis of belief (Dube et al., 2010; Trippas et al., 2013). As such, we assess this question by using Receiver Operating Characteristics (ROC) curve, which models sensitivity and bias separately.

2.3.2. ROC analysis

ROC curve analyses were conducted using the ROC Toolbox (Koen, Barrett, Harlow, & Yonelinas, 2017) on the data to test the effects of believability and language on accuracy and response bias independently. ROCs model proportion of hits (responding 'valid' when the conclusion is indeed valid) and false alarms (responding 'valid' when the conclusion is invalid) for each confidence level. These

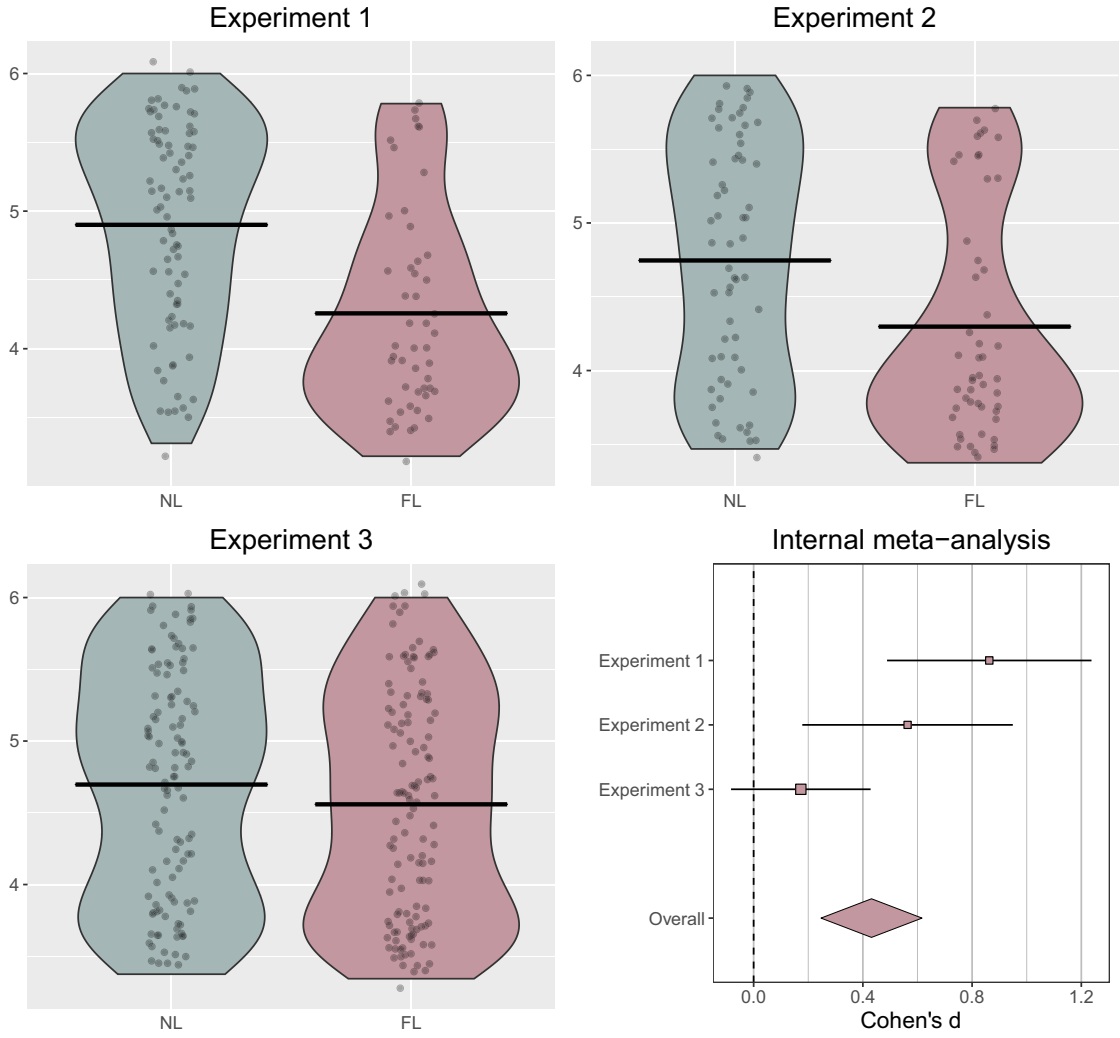

Fig. 1. Accuracy (weighted by confidence) for Experiments 1-3 and a forest plot of a meta-analytic effect size. Individual points are participant means, and group mean are indicated by horizontal bar.

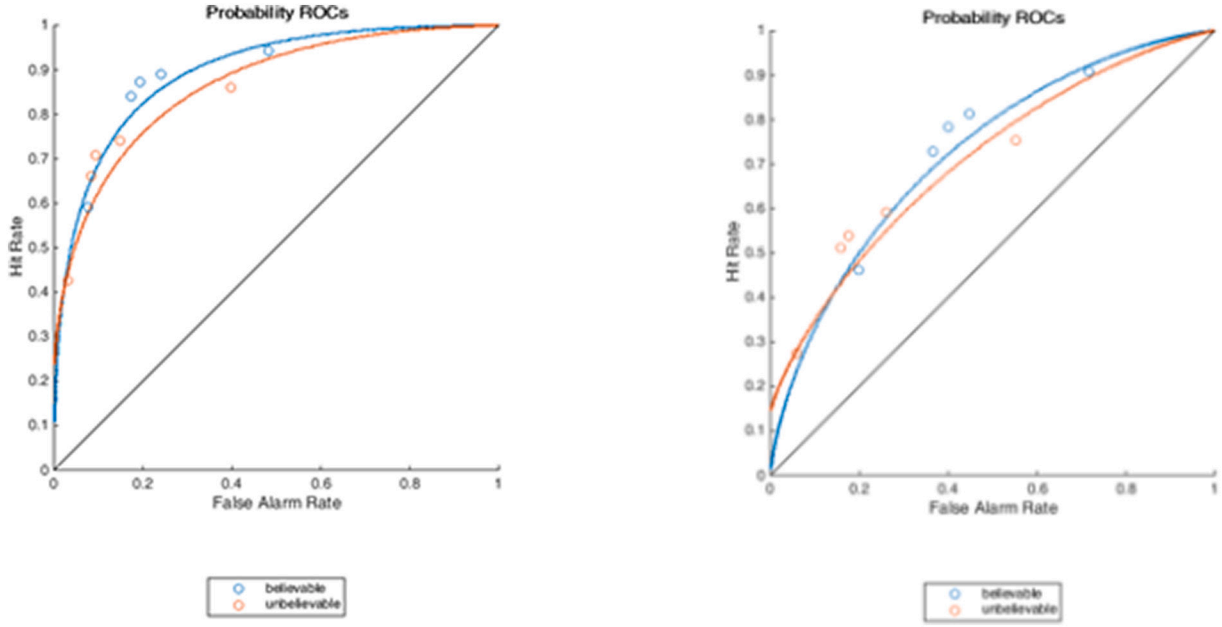

Fig. 2. ROC's for reasoning in NL (left), and FL (right) in Experiment 1. Positioning of the points on the curve represents the response bias (points toward the top-right corner suggest liberal response bias, greater willingness to endorse conclusions regardless of their logical status). Area under the curve corresponds to the accuracy in distinguishing between valid and invalid conclusions.

proportions are plotted as a curve, and the ROC Toolbox will provide several parameters indexing accuracy and response bias. Accuracy refers to the ability to correctly distinguish valid from invalid syllogisms and is usually a product of Type II processing. Response bias refers to a responding strategy, irrespective of actual logical validity and is a product of Type I processing. Liberal response bias involves an overall tendency (independent of logical structure or accuracy) to provide responses of "valid" as opposed to responses of "invalid". Conservative response bias, involves an overall tendency to provide responses of "invalid".[6]

We briefly explain how to interpret the ROC curve. As a point on the ROC curve approaches the top-left corner, where the Hit Rate is 1 and the False Alarm Rate is 0, it approaches maximum accuracy. As a point on the ROC curve approaches the top-right corner (where both hits and false alarms are 1, indicating the participant is always responding 'valid'), it approaches the most liberal response bias. ROC curves are fit to the points themselves, and as such an ROC curve that has greater area below, or toward the bottom right, represents greater accuracy, and an ROC whose points are plotted close to the top right represents more liberal bias. This type of analysis allows us to disentangle participants' tendency to overall endorse syllogisms (captured by the response bias parameter β, and the aggregate decision criterion *c*) from their overall accuracy (captured by area under the ROC curve, or AUC).

We test whether participants' accuracy in reasoning was different across language conditions (a main effect of language on AUC; a replication of the results of the traditional analysis) and whether believability affected response bias in each of our language conditions (i.e., whether a belief bias is present). Finally, we test whether response bias, the tendency to respond 'valid' regardless of logical validity, differs between believable and unbelievable syllogisms more for NL or FL reasoners (a language by believability interaction).

We used the ROC Toolbox to fit the dual-process signal detection (DPSD) model to our data, tabulated as frequencies within each confidence bin. We did this separately for participants in the NL and FL conditions, and separately for believable and unbelievable syllogisms within each group (Fig. 2).

A G-test indicated that the DPSD model fit both the NL and FL data well, both $Gs \geq 37.05$, $ps < .001$. Accuracy did not meaningfully differ between believable and unbelievable syllogisms in either NL ($AUC_B = 0.875$, $AUC_U = 0.842$) nor FL ($AUC_B = 0.711$, $AUC_U = 0.693$) reasoners. This is consistent with claims by Dube et al. (2010) suggesting that believability of a conclusion does not affect accuracy. However, accuracy is considerably greater in the NL (average AUC = 0.859) compared to the FL condition (average AUC = 0.702).[7] There was no apparent believability by language interaction. This analysis shows that people were considerably less accurate when reasoning in their FL. It also shows that reasoners' accuracy is similarly affected by the believability of the syllogism in either language.

For NL reasoners, response bias parameter β differed descriptively between believable and unbelievable syllogisms ($\beta_B = 0.76, \beta_U = 2.05$), and aggregate decision criterion *c* also differed by our inferential approach (detailed in footnote 7; $c_B = -0.134$, $c_U = 0.386$), $t(8) = 4.03$, $p = .004$. That is, participants' response criterion was significantly more liberal for believable relative to unbelievable syllogisms. Similar effects of believability were observed for FL reasoners, whose response bias parameter β differed descriptively between believable and unbelievable syllogisms ($\beta_B = 0.76$, $\beta_U = 1.53$), and aggregate decision criterion *c* also differed by our inferential approach between believable and unbelievable syllogisms ($c_B = -0.265$, $c_U = 0.417$), $t(8) = 5.55$, $p < .001$. That is, belief bias is present in both language conditions. The difference in response bias – between believable and unbelievable –

[6] To further clarify why one would adopt a strategy to err in a particular way (false alarms over miss) consider an oncologist, who would rather treat a healthy person (false alarm) over missing a sick patient and letting this person die (miss). Here, the oncologist adopted a liberal response bias. In this experiment, such strategy is captured in ROC as response bias with negative parameter c, or low parameter β.

[7] The ROC Toolbox does not provide a standard error or other measure of spread for measures of accuracy, so these comparisons are descriptive. However, we can reasonably infer some degree of spread for our measures of response bias because they are approximated with a standard error by the ROC Toolbox. For these analyses, we compare aggregate decision criterion *c* using the largest (and thus most conservative) of the individual bootstrap standard errors for the individual criteria ($c1 \ldots c5$) in the comparison.

Table 1
Conflict detection indices in Experiments 1–3.

Experiment		Native language			Foreign language		
		Congruent correct	Incongruent correct	Incongruent incorrect	Congruent correct	Incongruent correct	Incongruent incorrect
1	*n*	80	80	68	49	48	48
	Confidence	2.61 (0.41)	2.50 (0.44)	2.38 (0.49)	2.43 (0.41)	2.35 (0.43)	2.37 (0.49)
	Reaction times (log)	1.25 (0.11)	1.29 (0.11)	1.32 (0.20)	1.24 (0.14)	1.26 (0.14)	1.29 (0.23)
2	*n*	61	61	51	50	50	45
	Confidence	2.71 (0.29)	2.54 (0.25)	2.47 (0.45)	2.49 (0.38)	2.39 (0.46)	2.41 (0.47)
	Reaction times (log)	1.18 (0.16)	1.23 (0.16)	1.24 (0.26)	1.16 (0.14)	1.16 (0.16)	1.16 (0.26)
3	*n*	115	115	103	124	124	108
	Confidence	2.71 (0.28)	2.57 (0.35)	2.49 (0.45)	2.52 (0.42)	2.44 (0.44)	2.33 (0.49)
	Reaction times (log)	1.14 (0.16)	1.18 (0.19)	1.20 (0.25)	1.16 (0.17)	1.20 (0.20)	1.21 (0.24)

Note: Data presented as mean (standard deviation). No data is presented for incorrect congruent trials, because they are clearly erroneous, and very rare.

(i.e., the degree of belief bias) did not differ by our inferential approach across NL and FL ($\Delta c_{FL} = 0.683$, $\Delta c_{NL} = 0.520$), $t(8) = 1.26$, $p = .243$. That is, believability has no less of a biasing effect on those reasoning in their FL as compared to their NL.

Results from our ROC analysis suggest that FL reasoners are, compared to NL reasoners, overall much less accurate but no less biased by believability. That is, their validity judgments are *no less* likely than NL reasoners' to be made on the basis of believability but (at least for conflict problems) *far less* likely than NL reasoners' to be made on the basis of logic.

2.3.3. Conflict detection analysis

In the analyses reported above, we established that participants in their FL were less accurate in logical reasoning. This could be because people in their FL either fail to successfully complete their Type II processing, or perhaps because they fail to even engage Type II processing at all. In other words, reasoning in one's FL can either distort the rule-based Type II processing (i.e., logical validity), or amplify the biased Type I processing (i.e., believability).

To identify the mechanism responsible for decreased accuracy of FL reasoners, we compared the confidence and reaction times for correct responses on congruent trials (wherein belief and logic cue the same response) to incorrect responses on incongruent trials (wherein belief and logic cue opposite responses). If no Type II processing is engaged, a person's judgement will be intuitive and predominantly driven by judgement of believability of the conclusion rather than by the logical structure of the syllogism. Differently put, answering according to beliefs in these types of syllogisms usually signifies lack of reflection.

The critical difference between congruent and incongruent syllogisms used in this research is that the presumably ignored dimension of the syllogism (i.e., logical validity) is congruent with the dominant intuition (i.e., believability) in one example and incongruent in the other. If the logical validity is somehow processed, even when it does not translate to the ultimate response, it would conflict with the selected response, slow down response times, and decrease confidence only in incongruent syllogisms but not in congruent syllogisms. If, however, an intuition about the logical structure of the syllogism is simply not produced, it cannot interfere with the belief-based intuition. In that case, there should be no difference between reaction times and confidence ratings in congruent vs. incongruent syllogisms.

For completeness we also report the differences between correctly answered congruent syllogisms and correctly answered incongruent trials. Obviously, engaging in reflection to override intuitions should take longer than answering based on intuition only. Hence, it is expected that such trials take longer than congruent correct trials. However, no specific predictions regarding confidence can be derived from the theories of reasoning, since to answer correctly an incongruent syllogism one usually has to detect the conflict, and override it with reflection. In such cases, one can be less confident because of the conflict detection helped one to realize the task is difficult. However, confidence might subsequently increase because after successfully employing Type II processing one believes their response is now correct.

Table 1 presents conflict detection indices across language conditions and trial types. We analyzed confidence ratings and reaction times, but the former parameter seems to be a more faithful, and less noisy, index of conflict detection (Bago & De Neys, 2020).

For confidence ratings, we observed that participants were indeed more confident in congruent trials they answered correctly as compared to incongruent trials they answered correctly, $F(1, 126) = 16.31$, $p < .001$, $\eta_P^2 = 0.115$, and compared to incongruent trials they answered incorrectly, $F(1, 114) = 12.83$, $p = .001$, $\eta_P^2 = 0.101$. This suggests that people who processed incongruent trials were indeed less confident in their responses, even if their ultimate response was incorrect because it relied on the intuitive believability assessment.

When comparing confidence in incorrect incongruent responses to confidence in correct congruent responses, we also found a statistically significant language by congruence interaction, $F(1, 114) = 4.70$, $p = .032$, $\eta_P^2 = 0.040$, but no statistically significant main effect of language, $F < 1$. Decomposing the interaction with simple effects, using the Sidak correction for multiple comparisons, we see that, in NL participants, confidence was lower for incorrect-incongruent trials compared to correct-congruent trials, $F(1,114) = 19.97$, $p < .001$, $\eta_P^2 = 0.149$. No such difference was observed in FL participants, $F < 1$. When comparing confidence in correct incongruent responses to the confidence in correct congruent responses, no language by congruence interaction was observed for incongruent correct trials vs congruent correct trials, $F(1,126) = 1.31$, $p = .255$, $\eta_P^2 = 0.010$. Eyeballing Table 1 suggests, however, that conflict detection (reflected in lower confidence in incongruent trials) was stronger in NL. To wrap up, this analysis suggests that participants who erroneously responded on the basis of belief were less confident in their response, relative to correctly responding on the basis of belief, but only in their native language. No such difference in confidence was observed in their foreign language. This, then, suggests that no conflict detection occurred in FL.

To analyze reaction times, we first log-transformed them so that their distribution would be normal and less affected by possible outliers. Consistent with the findings vis-à-vis confidence, participants tended to respond faster in congruent correct trials compared to incongruent correct, $F(1, 126) = 11.65$, $p = .001$, $\eta_P^2 = 0.085$, and to incongruent incorrect trials, $F(1, 114) = 3.26$, $p = .074$, $\eta_P^2 = 0.028$, but this latter difference was not statistically significant. When comparing correct congruent to incorrect incongruent trials, there was no statistically significant main effect of language $F(1, 114) = 2.23$, $p = .138$, $\eta_P^2 = 0.019$, but again a significant language by congruence interaction, $F(1, 114) = 4.35$, $p = .039$, $\eta_P^2 = 0.037$. Decomposing the interaction with simple effects, using the Sidak correction for multiple comparisons, we see that conflict was detected in the NL condition, $F(1,114) = 9.14$, $p = .003$, $\eta_P^2 = 0.074$, but not in the FL condition, $F < 1$. No language by congruence interaction was observed when

comparing correct congruent to correct incongruent trials, $F(1, 126) = 2.43$, $p = .121$, $\eta_P^2 = 0.019$.

Reasoners who solved incongruent syllogisms using their native language were less confident, and required more time to process them, compared to when solving congruent syllogisms. Reasoners who solved the syllogisms using their foreign language were just as confident, and required about the same time to process regardless of whether the syllogism was congruent or incongruent. Therefore, our results provide preliminary evidence for distorted conflict detection in FL reasoning. In other words, FL reasoners failed to accurately differentiate between congruent and incongruent syllogisms, decreasing the likelihood of engaging in Type II processing in incongruent syllogisms when it is required to override an incorrect intuitive response. In turn, FL reasoners responded more often with said incorrect, intuitive response. There is also another potential source of error due to failed conflict detection. Congruent trials are designed so that they are valid every time the conclusion is believable, and invalid every time the conclusion is unbelievable. Hence, responding in line with belief is guaranteed to be correct. Since FL reasoners failed to detect that a congruent syllogism is in fact congruent, it can lead them to unnecessarily reflect in congruent trials, risking error due to failed Type II processing. We argue that failed conflict detection is the mechanism responsible for lower accuracy as a result of reasoning in one's foreign language.

2.4. Discussion

Our findings suggest that reasoners in their FL are less accurate (i.e., they produce fewer correct responses) and no less biased (i.e., they rely on believability of a conclusion at least as much) in their second language. Our results suggest that this is because FL reasoners more often fail to trigger Type II processes to override intuitions when logical validity is incongruent with believability. Our findings cannot be explained by FL working as cognitive load and overloading working memory required for reflection. This is because conflict detection that is normally robust to imposing cognitive load (Białek & De Neys, 2017; W. De Neys, 2006; Franssens & De Neys, 2009) was distorted in FL reasoners. In other words, FL reasoners likely failed to detect instances when Type II processing is required, while cognitive load would likely only prevent reasoners from completing Type II processing.

The observed decrease in reasoning accuracy in one's FL is surprising given the seeming robustness of the FLE within the decision-making literature (but see Białek et al., 2020; Hayakawa et al., 2019; Mækelæ & Pfuhl, 2019; Muda, Walker, et al., 2020, Vives et al. for recent failures to find the FLE in decision making). One possible explanation for this discrepancy is our use of pseudo-words in our premises. It is possible that these words made participants feel as if they did not understand the task. This perceived lack of understanding may have altered the approach taken by participants in the task. For example, participants may have been discouraged from engaging Type II processing. Another issue is a relatively large drop-out based on self-reported understanding in NL. It is unclear why a participant in their native language would report not understanding the task, and the most plausible explanation is that participants may have misinterpreted the question about understanding and responded to this question with an assessment of their performance (e.g., “I don't think I did very well, I must not have understood the task”).

To address these issues, we conducted Experiment 2 wherein we replaced meaningless words with graphical symbols to increase the understandability of the syllogisms.

3. Experiment 2

3.1. Participants

Ultimately, we analyzed data from 111 participants. We recruited 228 participants ($n = 151$ female, $M_{AGE} = 20$, $SD = 0.87$[8]) from UMCS University in Lublin in exchange for $5 compensation. We dropped the data from participants who reported understanding of the materials to be lower than 5 on a 10-point scale ($n = 38$ in NL condition, and $n = 38$ in FL condition), and a further $n = 41$ whose accuracy was below chance.

3.2. Materials and procedure

The procedure was identical to Experiment 1, with one significant alteration. Here, we tried to reduce the potential confound in understanding the premises by replacing abstract, meaningless words with abstract, geometric shapes as follows:

No △ are trees
All maples are △
Therefore, some trees are maples

We hoped that, with this change, fewer people would think the middle term of the syllogism (e.g., the △ in the example above) had a true meaning they simply fail to understand, thanks to which processing the syllogisms in their entirety would be easier.

3.3. Results

3.3.1. Traditional analysis

As in Experiment 1, we found that accuracy in the NL condition was higher than in the FL condition, $t(109) = 2.96$, $p = .004$, $d = 0.56$, 95% CI [0.18, 0.95] (Fig. 1, Panel B). As in Experiment 1, there was no difference in response times between NL and FL reasoners, $t(109) = 1.28$, $p = .204$, $d = 0.24$, 95% CI [−0.13, 0.62].

3.3.2. ROC analysis

A G-test indicated that the DPSD model fit both the foreign language and native language data well, both $Gs \geq 27.24$, $ps \leq .001$ (Fig. 3). There were descriptively small differences in accuracy between believable and unbelievable for both NL ($AUC_B = 0.824$, $AUC_U = 0.795$) and FL ($AUC_B = 0.733$, $AUC_U = 0.691$) reasoners. There was, again, no believability by language interaction. This analysis demonstrates that believability of a conclusion does not have a large effect on reasoning accuracy in either language. Consistent with Experiment 1, NL reasoning (average AUC = 0.809) was descriptively more accurate than FL reasoning (average AUC = 0.712).

For NL reasoners, response bias parameter β differed descriptively between believable and unbelievable syllogisms ($\beta_B = 0.68$, $\beta_U = 2.18$), and aggregate decision criterion c also differed by our inferential approach ($c_B = -0.235$, $c_U = 0.468$), $t(8) = 4.53$, $p = .002$. As in Experiment 1, NL reasoners adopted more liberal response criterion for believable conclusions, demonstrating belief bias (Dube et al., 2010). In terms of response bias for FL reasoners, β ($\beta_B = 0.88$, $\beta_U = 1.46$) and c ($c_B = -0.110$, $c_U = 0.392$) differed between believable and unbelievable syllogisms, $t(8) = 3.27$, $p = .011$. Both NL and FL reasoners showed belief bias. As in Experiment 1, the difference in response bias between believable and unbelievable syllogisms (i.e., the belief bias effect) does not differ across NL and FL reasoners ($\Delta c_{FL} = 0.502$, $\Delta c_{NL} = 0.702$), $t(8) = 1.77$, $p = .115$. Once again, FL show a belief bias that is similar in strength to that of NL reasoners and are overall far less accurate than NL reasoners. In other words, NL and FL reasoners were equally affected by the believability of the conclusion, which suggests they comprehended the syllogisms equally correctly.

[8] Because of a coding error, demographic information cannot be tied to individual participants' data, so the demographic information presented here is for the total sample before exclusions.

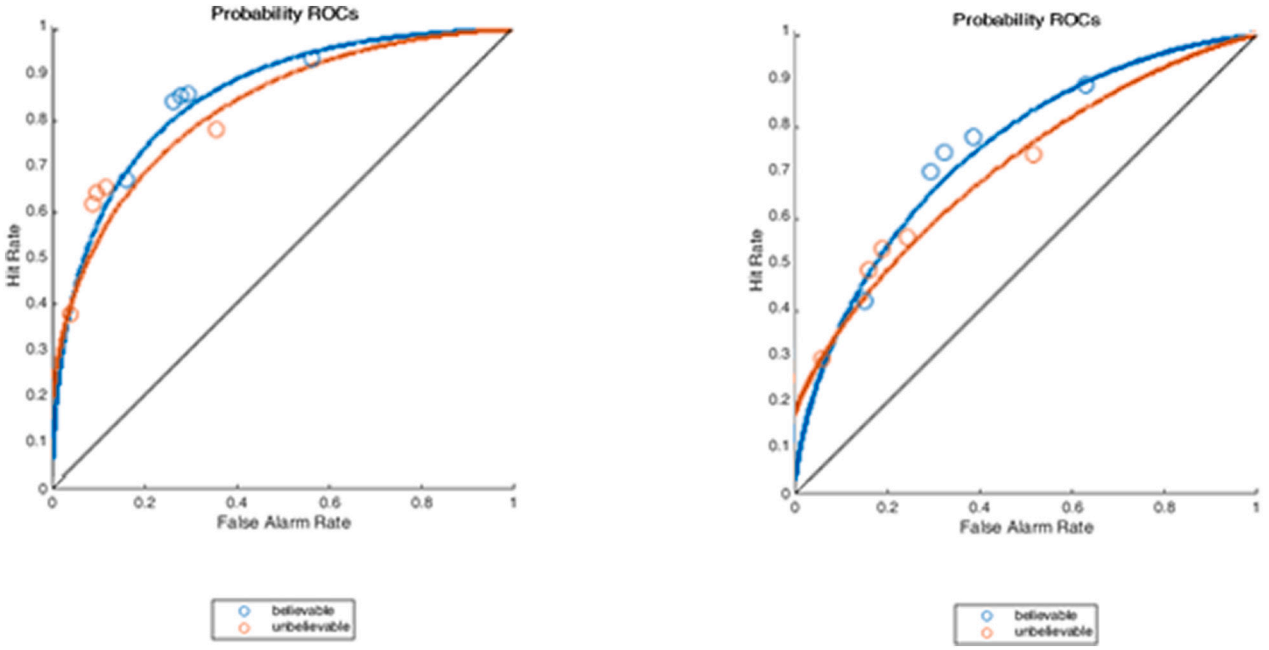

Fig. 3. ROC's for reasoning in NL (left), and FL (right) in Experiment 2. Positioning of the points represents the response bias (points toward the top-right corner suggest greater willingness to endorse conclusions regardless of their logical status). Area under the curve corresponds to the accuracy in distinguishing between valid and invalid conclusions.

3.3.3. Conflict detection analysis

Consistent with Experiment 1, confidence ratings were higher for correct-congruent trials than incorrect-incongruent trials in NL, $F(1,50) = 14.64$, $p < .001$, $\eta_P^2 = 0.227$, suggesting that conflict was detected (Table 1). There was no difference in confidence ratings between these trial types in the FL condition, $F(1,44) < 1$. However, in contrast to Experiment 1, the interaction was statistically non-significant, $F(1,94) = 2.84$, $p = .096$, $\eta_P^2 = 0.029$. For response times, we found a trend suggesting conflict detection, i.e., longer decision times for incorrect-incongruent trials compared to correct-congruent trials, in NL, $F(1,50) = 6.78$, $p = .012$, $\eta_P^2 = 0.119$, but not in FL, $F(1, 44) < 1$. Again, the congruence by language interaction was non-significant, $F(1, 94) = 2.36$, $p = .128$, $\eta_P^2 = 0.024$.

For completeness we report effects on confidence and reaction times between correct-congruent and correct-incongruent trials. For confidence, participants were more confident in correct-congruent trials, $F(1, 109) = 25.84$, $p < .001$, $\eta_P^2 = 0.192$, and when reasoning in their native language, $F(1, 109) = 8.56$, $p = .004$, $\eta_P^2 = 0.073$. We observed no language by congruence interaction, $F(1, 109) = 1.66$, $p = .201$, $\eta_P^2 = 0.015$.

For reaction times, we observed longer reaction times for incongruent trials, $F(1, 109) = 6.17$, $p = .015$, $\eta_P^2 = 0.054$, no effect of language, $F(1, 109) = 2.67$, $p = .105$, $\eta_P^2 = 0.024$, and a language by congruence interaction, $F(1, 109) = 6.37$, $p = .013$, $\eta_P^2 = 0.055$. The interaction occurred because participants were slower in incongruent trials in their NL, $F(1, 60) = 13.99$, $p < .001$, $\eta_P^2 = 0.189$, but no such difference was observed in FL, $F < 1$. In general, in trials they got correct, our participants were faster and more confident in congruent trials.

3.4. Discussion

The data seems to be descriptively consistent with data from Experiment 1. However, we found relatively weaker evidence of thinking in FL distorting conflict detection. We have seen that our participants were less confident in incongruent trials, and required longer reaction times, but only in their NL. In their FL, no such differences have been observed. What can be inferred, compared to experiment 1, is that using foreign language does not knock out the conflict detection, but only weakens it. Obviously, in each experiment FL participants are still engaging Type II processes to some degree. Otherwise, their accuracy would be at 50%.

Note that because of high attrition, both non-significant interaction tests might be underpowered. As a possible solution to this, we decided to run a third, preregistered experiment. This time we changed the data reduction policy to be less strict (i.e., asked about self-rated proficiency instead of understanding, see Keysar et al., 2012; Geipel et al., 2015; Muda, Pieńkosz, Francis, & Białek, 2020 for similar selection criterion), and further increased our sample size.

4. Experiment 3

4.1. Participants

This experiment was preregistered at AsPredicted.org https://aspredicted.org/blind.php?x=pd6h7r. Ultimately, we analyzed data from 238 participants. We recruited 305 participants ($n = 186$ female, $M_{AGE} = 22.3$, $SD = 2.86$) from UMCS University in Lublin in exchange for $5 compensation. We dropped the data from participants who reported their English proficiency to be lower than 5 on a 10-point scale ($n = 10$ in NL condition, and $n = 15$ in FL condition), and a further $n = 42$ whose accuracy was below chance.

4.2. Materials and procedure

The procedure was identical to Experiment 1 and 2, with one significant alteration. Here, we tried to reduce the potential confound in understanding the premises even further, by replacing abstract, meaningless words (Experiment 1) or geometric shapes (Experiment 2) with letters, as follows:

No P are trees
All maples are P
Therefore, some trees are maples

We hoped that, with this change, processing the syllogisms in their

entirety would be easier. Such structure resembles the one used in logic classes, which all Polish students attended in high school, and in their first-year undergraduate classes.

4.3. Results

4.3.1. Traditional analysis

As in Experiments 1 and 2, we found that accuracy was higher in the NL condition, but this time the difference was non-significant, $t(236) = 1.25$, $p = .106$ (one-tailed), $d = 0.16$, 95% CI [−0.09, 0.42] (Fig. 1, Panel C). As in the previous experiments, there was no difference in response times between NL and FL reasoners, $t(236) = 1.09$, $p = .279$, $d = 0.14$, 95% CI [−0.11, 0.40].

4.3.2. ROC analysis

A G-test indicated that the DPSD model fit both the foreign language and native language data well, both $Gs \geq 67.23$, $ps \leq .001$ (Fig. 4). There were descriptively small differences in accuracy between believable and unbelievable for both NL ($AUC_B = 0.875$, $AUC_U = 0.842$) and FL ($AUC_B = 0.804$, $AUC_U = 0.757$) reasoners. There was, again, no believability by language interaction. This analysis demonstrates that believability of a conclusion does not have a large effect on reasoning accuracy in either language. Consistent with Experiments 1 and 2, NL reasoning (average AUC = 0.859) was descriptively more accurate than FL reasoning (average AUC = 0.780).

For NL reasoners, response bias parameter β differed descriptively between believable and unbelievable syllogisms ($\beta_B = 0.764$, $\beta_U = 2.05$), and aggregate decision criterion c also differed by our inferential approach ($c_B = -0.134$, $c_U = 0.386$), $t(8) = 4.53$, $p = .028$. In terms of response bias for FL reasoners, β ($\beta_B = 0.87$, $\beta_U = 1.68$) and c ($c_B = -0.099$, $c_U = 0.395$) differed between believable and unbelievable syllogisms, $t(8) = 20.82$, $p < .001$. Again, the difference in response bias between believable and unbelievable syllogisms (i.e., the belief bias effect) does not differ across NL and FL reasoners ($\Delta c_{FL} = 0.494$, $\Delta c_{NL} = 0.520$), $t(8) = 0.20$, $p = .846$. Once again, FL reasoners show a belief bias that is equivalent to that of NL reasoners and are overall far less accurate than NL reasoners.

4.3.3. Conflict detection analysis

Table 1 presents the descriptive statistics for all conflict detection indices. We first compared the critical correct-congruent and incorrect-incongruent trials. To remind you, both are assumed to be responded based on the believability dimension, with the difference being that, in incongruent trials, ignored validity cued an opposite response. If one responds along the belief dimension and did not even process validity, one should be just as confident in both types of trial. If, however, one processed validity in the background, one should be less confident in their decision because of the belief-validity conflict. If a difference is found, it evidences some degree of conflict detection, a process critical to triggering Type II processing (De Neys & Pennycook, 2019).

For confidence ratings, congruent-correct trials were answered with greater confidence than incongruent-incorrect, $F(1, 208) = 47.73$, $p < .001$, $\eta_P^2 = 0.186$. Moreover, FL participants responded with lower confidence than did NL participants, $F(1, 209) = 13.10$, $p < .001$, $\eta_P^2 = 0.060$. We observed no language by congruence interaction, $F < 1$. For reaction times, participants responded faster in congruent trials, $F(1, 208) = 31.35$, $p < .001$, $\eta_P^2 = 0.131$, with no other effects being significant F's < 1. These results suggest that conflict was detected in both languages with similar strength.

For completeness, we report comparisons between congruent-correct and incongruent-correct trials. Regarding the confidence ratings, participants were more confident in congruent trials, $F(1, 236) = 46.89$, $p < .001$, $\eta_P^2 = 0.166$, and in their native language, $F(1, 236) = 12.87$, $p < .001$, $\eta_P^2 = 0.052$. these effects were qualified by a language by congruence interaction, $F(1, 236) = 4.86$, $p = .028$, $\eta_P^2 = 0.020$. The interaction resulted because the difference between confidence in congruent and incongruent trials was greater in NL, $F(1, 114) = 38.00$, $p < .001$, $\eta_P^2 = 0.250$ than in FL, $F(1,122) = 11.62$, $p = .001$, $\eta_P^2 = 0.087$. Regarding reaction times, we found only a main effect of congruence, $F(1, 236) = 34.35$, $p < .001$, $\eta_P^2 = 0.127$, with the effect of language, $F(1, 236) = 1.11$, $p = .293$, $\eta_P^2 = 0.005$, and language by congruence interaction, $F < 1$, both not being significant. Hence, individuals who successfully overrode belief intuitions were less confident in their response compared to congruent trials, but this decrease was greater in NL compared to in FL.

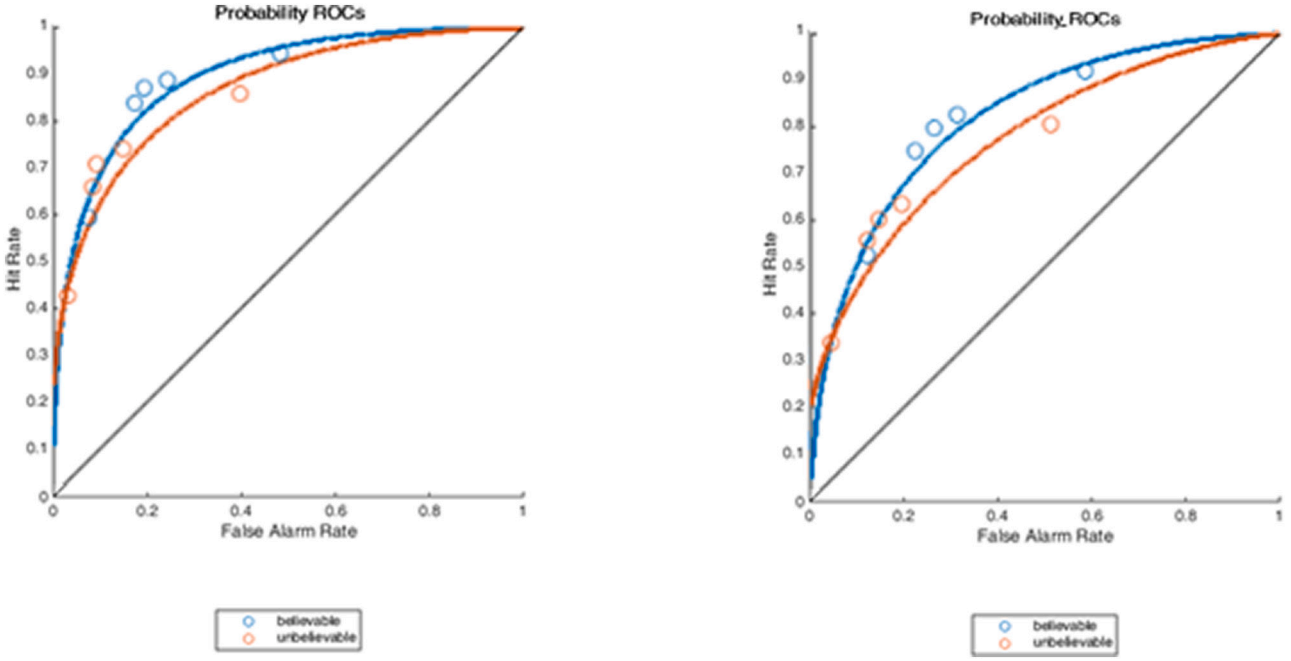

Fig. 4. ROC's for reasoning in NL (left), and FL (right) in Experiment 3. Positioning of the points represents the response bias (points toward the top-right corner suggest greater willingness to endorse conclusions regardless of their logical status). Area under the curve corresponds to the accuracy in distinguishing between valid and invalid conclusions.

4.4. Discussion

We found descriptively similar patterns of results as in Experiments 1 and 2: FL reasoners were characterized by lower reasoning accuracy caused by lower sensitivity to logical status, and weaker conflict detection. However, it is worth noting that the results of Experiment 3 fell short of statistical significance in several places, despite being generally consistent in direction with those of the first two experiments. This could be the case for one of multiple reasons. Perhaps, across three experiments we have captured the exact same effect and any differences between experiments are due to random chance or issues with power. Indeed, a clear pattern of only significant multi-study results has been often deemed too good to be true (Francis, 2014; Schimmack, 2012), and suspicious on those grounds. Additionally, we have not powered our studies to test an attenuated interaction,[9] i.e., of the sort in which one groups shows an effect of interest, and the other group shows same effect, but weaker. This is because in Experiment 1 we observed no conflict detection in FL, and also because attenuated interaction tests require larger sample sizes than are plausible to collect (Giner-Sorolla, 2018; Simonsohn, 2014).

The other possibility is that there are causal differences between the types of syllogism we have used across three experiments. This possibility is worth entertaining. One can see from inspection of Fig. 1 that, if differences between experiments exist, they are primarily in the FL condition. As such, these differences are likely attributable to language, and at least by the metrics of interest unique to the present investigation. Imagine, then, that we have sampled three times from the hypothetical distribution of types of syllogism. These types can be categorized by dependence on language, from language-rich to language-poor. If language is a causal factor as we have hypothesized here, one would reasonably expect the effects to be strongest in Experiment 1, where the syllogisms most closely mirror language by implanting nonsense words in the premises of each syllogism. As we move away from traditional language, as in Experiments 2 and 3, one might predict that the effects of language decrease. In Experiment 3, clauses like: “All trout are P”, and “All P are fish”, resemble the notation typically used in logic classes. It should be noted that Polish students all have such classes both at high-school and university levels. Our observed effects, across three experiments, follow exactly this pattern: a large FLE when the syllogisms most closely mirror language and a decreasing FLE as the syllogisms approach formal logic. As a final point on this topic, it is worth considering which type of problem is likely to yield real-world consequences. If left to choose between language-like syllogisms and those that emulate propositional logic, the former more closely mirrors real-world reasoning. That is, the observed effects are strongest where reality intersects the distribution of tasks tested here.

Because of some inconstancies in significances in the statistical tests, we decided to pool our data, and run two critical tests of reasoning (as encompassed by our problems collectively) in a foreign language: accuracy and conflict detection. This way we will be able to assess whether accuracy of reasoning deteriorates in foreign language across a broad range of problem types, and if yes, whether it can be attributed to decreased sensitivity to a conflict between believability and validity of a conclusion. This pooled analysis is the critical test of our hypotheses because it has the most power to detect any effects.

5. Pooled data analysis

We collapsed data from Experiments 1–3, allowing data from $n = 479$ participants to be analyzed. Fig. 1 panel D presents a forest plot of the effect sizes of a difference between reasoning accuracy in NL and in FL. A meta-analytic effect size is d = 0.43, 95% CI [0.25–0.61]. Also, an analysis of the pooled data from Experiments 1–3 shows that accuracy is significantly lower in FL ($M = 4.44$, $SD = 0.79$) than in NL ($M = 4.77$, $SD = 0.79$), $t(477) = -4.68$, $p < .001$, $d = 0.42$, 95% CI [0.24–0.60]. Having enough power to do so, we explored whether this decrease in reasoning accuracy is somewhat different for different types of syllogisms with regard to their believability and validity. To this end, we ran a 2(validity, within-subject) x 2(believability, within-subject) x 2(language, between-subject) ANOVA. We again found a robust effect of language, $F(1, 477) = 21.91$, $p < .001$, $\eta_p^2 = 0.044$, which did not interact with believability nor validity of a conclusion, both F's < 1. All these analyses provide a strong evidence that, contrary to what could be expected based on the literature review, bilinguals reason worse when using their foreign language.

In terms of the robustness check, we confirmed past findings on belief bias in which people were more accurate for valid vs. invalid, $F(1, 477) = 33.42$, $p < .001$, $\eta_p^2 = 0.065$, and for unbelievable vs. believable syllogisms, $F(1, 477) = 81.37$, $p < .001$, $\eta_p^2 = 0.146$. These effects were qualified by a logic by belief interaction, $F(1, 477) = 253.14$, $p < .001$, $\eta_p^2 = 0.347$. All of these effects have been found in the past, and were successfully replicated here (Evans, Barston, & Pollard, 1983; Klauer, Musch, & Naumer, 2000).

Next, we investigate conflict detection indices across languages (Fig. 5). When comparing correct-congruent to incorrect-incongruent trials, we observed strong evidence of conflict detection in confidence ratings ($M_{congruent} = 2.57$, $SD = 0.39$; $M_{incongruent} = 2.41$, $SD = 0.48$), $F(1,421) = 70.85$, $p < .001$, $\eta_p^2 = 0.144$; and in reaction times, ($M_{congruent} = 1.18$, $SD = 0.16$; $M_{incongruent} = 1.23$, $SD = 0.24$), $F(1,421) = 32.49$, $p < .001$, $\eta_p^2 = 0.072$. We also found evidence for a language by congruity interaction, statistically significant with confidence as a dependent variable, $F(1,421) = 5.77$, $p = .017$, $\eta_p^2 = 0.014$, and trending with reaction times as a dependent variable $F(1,421) = 2.74$, $p = .099$, $\eta_p^2 = 0.006$. The interaction tests indicate conflict was detected in NL, $F(1,221) = 62.65$, $p < .001$, $\eta_p^2 = 0.221$ for confidence, and $F(1,221) = 27.79$, $p < .001$, $\eta_p^2 = 0.112$, for reaction times. Conflict detection was substantially weaker in FL, $F(1,200) = 19.92$, $p < .001$, $\eta_p^2 = 0.078$ for confidence, and $F(1,200) = 8.00$, $p = .005$, $\eta_p^2 = 0.038$ for reaction times, respectively. In other words, reasoners who responded along the belief dimension processed the logical dimension (and thus experienced greater conflict) more in their native language than they did in their foreign language.

A similar pattern is observed when comparing correct-congruent to correct-incongruent trials. We observed general evidence of conflict detection in confidence ratings ($M_{congruent} = 2.59$, $SD = 0.38$; $M_{incongruent} = 2.48$, $SD = 0.41$), $F(1,476) = 89.36$, $p < .001$, $\eta_p^2 = 0.158$; and in reaction times, ($M_{congruent} = 1.18$, $SD = 0.16$; $M_{incongruent} = 1.22$, $SD = 0.17$), $F(1,476) = 52.82$, $p < .001$, $\eta_p^2 = 0.100$. We also found evidence for a language by congruence interaction in confidence, $F(1,476) = 7.41$, $p = .007$, $\eta_p^2 = 0.015$, and in reaction times $F(1,476) = 4.87$, $p = .028$, $\eta_p^2 = 0.010$. The interaction tests indicate that conflict was detected in NL, $F(1,255) = 83.43$, $p < .001$, $\eta_p^2 = 0.247$ for confidence, and $F(1,255) = 49.33$, $p < .001$, $\eta_p^2 = 0.162$, for reaction times, but to a lesser extent in FL, $F(1,221) = 20.14$, $p < .001$, $\eta_p^2 = 0.084$ for confidence, and $F(1,221) = 11.68$, $p = .001$, $\eta_p^2 = 0.050$ for reaction times, respectively. In other words, reasoners who responded along the logical dimension (and successfully overrode their belief-based intuitions) were more conflicted in their native language compared to their foreign language.

As mentioned in previous sections, a potential problem with our data was that in Experiments 1 and 2 we rejected a significant number of participants. This could have biased the results. For example, we could have analyzed the data from participants who are qualitatively different from the ones rejected (i.e., more attentive, with higher cognitive abilities, or who default to a different type of processing). We explored our data retaining the entire sample of participants from all three experiments ($n = 744$). We found almost identical results as the ones reported above, with language by congruence interactions

[9] One of the estimates is that such tests require up to 16 times larger samples than those required to detect the main effect (Giner-Sorolla, 2018).

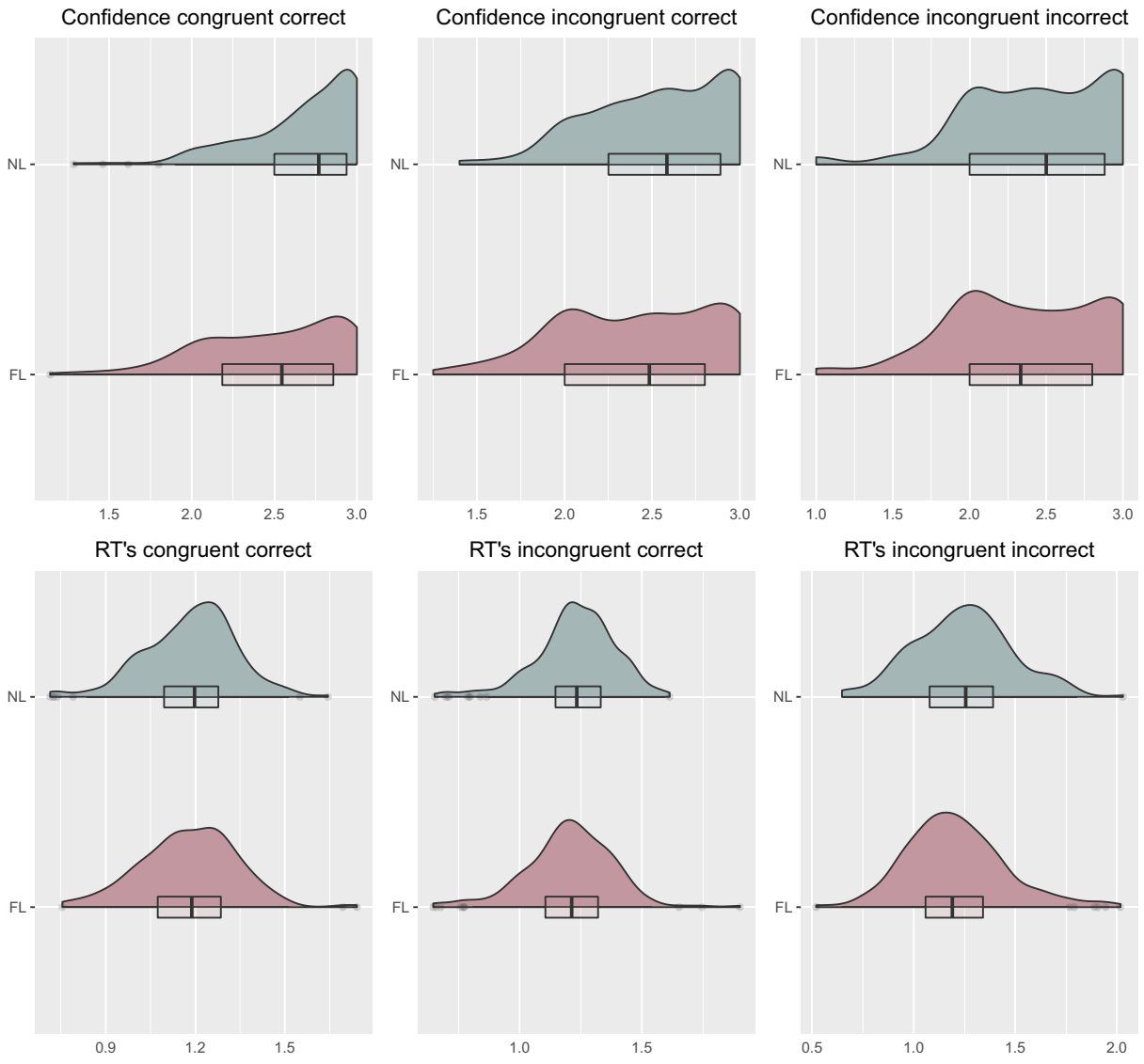

Fig. 5. Conflict detection indices for confidence and reaction times. Pooled data from Experiments 1-3.

significant for reaction times and confidence. Details of this analysis can be found in the Supplementary materials.

6. General discussion

Across three experiments, we found evidence for decreased accuracy in reasoning for participants reasoning in their foreign language. This effect can be attributed to a decrease in sensitivity to the logical structure of the syllogisms. That is, FL reasoners were less able than NL reasoners to distinguish logically valid from logically invalid syllogisms when asked to do so. FL reasoners were also not any less biased by the believability of the syllogisms than NL reasoners.

One avenue to consistently and correctly solve syllogisms is to engage reflective, Type II thinking in all circumstances. Utilizing this strategy incurs a large resource cost. Therefore, an alternative, efficient strategy would be to engage reflective, Type II processing only when it is necessary (i.e., when an intuitive, heuristic approach yields the incorrect answer) and to not engage reflective, Type II processing when it is unnecessary (i.e., when an intuitive, heuristic approach yields the correct answer). Explaining how one knows whether their intuitive output is correct without engaging in Type II processing is a major challenge for dual process theories (Evans & Stanovich, 2013; Pennycook et al., 2015). Recent consensus is that this strategy requires a metacognitive ability to discern when reflection is necessary and when it is not, that is, to detect conflict between belief-based and logic-based intuitions. Thanks to this intuitive conflict detection, people could engage in Type II processing more efficiently, without a substantial decrease in accuracy from unilateral engagement (De Neys & Pennycook, 2019; Handley & Trippas, 2015; Thompson et al., 2011). Most recent models of reasoning suggest that the initial response to a syllogism is accompanied by a Feeling of Rightness. When this Feeling of Rightness is low, suggesting the conflict between relevant intuitions is detected, it can trigger Type II processing (De Neys, 2014; Pennycook et al., 2015; Thompson et al., 2011). Our novel finding is that where NL reasoners are generally able to detect a conflict between competing intuitions and resolve it with Type II processing, FL reasoners are less sensitive to such conflict. In plain words, when two intuitive, Type I responses (in this case outputs representing logic and belief) conflict, NL reasoners more often than FL reasoners realize further reflection is required. In turn, NL reasoners allocate their cognitive effort more accurately, reflect more when it is required to correctly assess the status of a syllogism (and potentially less when reflection is not required) and, as a result, are more accurate in reasoning.

An alternative explanation why FL reasoners underperform in reasoning tasks is their Type II processing is distorted. We believe this explanation can be rejected. We have no evidence that there is a

difference in Type II reasoning quality between NL and FL reasoners. Evidence of this nature would be a difference in overall accuracy of reasoning (as tested in traditional analyses) or sensitivity to the logical structure of a syllogism (expressed as AUC in the ROC analyses) without commensurate differences in detection of belief-logic conflict. If this was observed, we could infer that reasoners detected the logic-belief conflict, but failed to resolve it with Type II processes. Since we observed a decrease in both reasoning accuracy and in conflict detection, we argue the crucial difference in accuracy between NL and FL reasoners is prior to the engagement of Type II processes. Further evidence for this position arises out of comparing Experiment 3 to the previous experiments. We observe both a decrease in the accuracy difference between NL and FL reasoners and also a decrease in the size of the conflict detection difference between NL and FL reasoners. If the difference between NL and FL reasoners is due to differences in ability to detect conflict, we would expect a change in conflict detection to inherently pull the reasoning accuracy in the same direction. This is exactly what we observe. Specifically, it appears NL reasoners are reasonably well calibrated as to when they should and should not engage Type II processing, but FL reasoners, generally speaking, are less able to identify situations in which engagement is beneficial. Where NL reasoners have a fairly accurate barometer for reflection, FL reasoners' barometer seems to be miscalibrated.

Above, we discussed our results from the perspective of the Dual-Process Theory. Let us now focus on the proposed explanations of the FLE. Previous studies about the FLE revealed that FL decision-makers are protected from several common heuristics and biases (Costa, Foucart, Hayakawa, et al., 2014; Gao et al., 2015; Hadjichristidis et al., 2019, Keysar et al., 2012), while some other research found no such effect (Białek et al., 2020; Mækelæ & Pfuhl, 2019; Muda, Walker, et al., 2020; Vives et al., 2018). The present research uses the reasoning domain to examine the FLE in a novel context. We see a reversal of the FLE in that FL reasoners are markedly worse in a syllogistic reasoning task. This provides us with a unique opportunity to scrutinize the mechanistic claims of the FLE. One potential mechanism responsible for this effect is that FL decision-makers might be more likely to engage in reflective, Type II processing. A potential competing mechanism is that FL decision-makers are simply less affected by the emotional, heuristic dimensions of commonly-used tasks. These explanations of the FLE allowed us to derive several predictions regarding reasoning. Costa, Foucart, Hayakawa, et al. (2014) and Hayakawa et al. (2016) hypothesized that, if the FLE is a result of cognitive differences between NL and FL reasoners, we should see an increase in accuracy for a task that requires Type II processing, like the one we have employed; however, if the FLE is a result of emotional differences, we should only observe differences as a result of reasoning in one's FL if the task is emotional in nature. Thus, in affect-poor tasks like syllogistic reasoning, we should see no effect of reasoning in a foreign language.

We observe neither of these outcomes. Instead, we see a decrease in overall accuracy for FL reasoners. This is predicted by neither of the aforementioned mechanisms. Where the emotional mechanism would predict no difference in accuracy in affect-poor syllogistic reasoning, we see a large difference. Where the cognitive mechanism would predict an increase in accuracy for FL reasoners, we find a decrease. Moreover, we observe substantially less detection of conflict (between two competing intuitions) for FL reasoners. This calls for a revision of the explanations of the FLE.

We propose a model of foreign language reasoning wherein the critical difference in a syllogistic reasoning task between FL and NL reasoners is the ability to detect conflict and allocate cognitive effort. There are several possible explanations in terms of the relative strength of FL reasoners' produced intuitions. Perhaps they have a higher threshold for conflict. This Conflict Threshold Model would predict that when experiencing equivalent conflict between competing intuitions, NL reasoners will detect a conflict and FL reasoners will not. This mechanism is not necessarily irreconcilable with the emotion-based mechanism (Costa, Foucart, Hayakawa, et al., 2014; Keysar et al., 2012; Vives et al., 2018). For example, De Neys, Moyens, and Ansteenwegen (2010) proposed that conflict detection is affective by nature. If it is the case that FL reasoners experience diminished emotion as a product of reasoning in their foreign language (Pavlenko, 2008, 2012), and the experience of emotion is crucial to the detection of conflict, it follows that they will be less able to detect conflict even when it is present. An alternative account might suggest reasoning in a foreign language stunts one's logical intuitions either to the point where they are non-existent, or simply to the point they are overwhelmed by belief-based intuitions. This Stunted Intuitions Model would predict that FL reasoners' intuitions might simply conflict less with one another. A problem with this model is a lack of a language effect on cognitive reflection test, in which the incorrect response is intuitive and the correct response is reflective (Białek et al., 2019; Costa, Foucart, Hayakawa, et al., 2014; Mækelæ & Pfuhl, 2019, but see Thompson, Pennycook, Trippas, & Evans, 2018 and Raoelison, Thompson, & De Neys, 2020 for evidence for intuitive origins of correct responses in the CRT). Finally, this model would predict increased belief bias in FL, which was not observed in this research.

Both of the mechanisms we propose are irreconcilable with the cognition-based account (Cipolletti, McFarlane, & Weissglass, 2016; Costa, Foucart, Hayakawa, et al., 2014; Keysar et al., 2012). It is difficult to imagine reasoning in a foreign language being both: 1) more influenced by Type II reflective processes, and 2) less accurate in syllogistic reasoning and less amenable to conflict detection.

One might be wondering how the Conflict Threshold and Stunted Intuitions Models proposed here are reconcilable with past findings showing benefits of reasoning in a foreign language. One avenue to understanding this seeming disjunction is to take the position that previous tasks showing a benefit of thinking in a foreign language used tasks that do not produce conflicting intuitions (e.g., in gambling the only available intuition is risk aversion, in framing the only available intuition reflects the decision frame). Perhaps the tasks wherein FL reasoners show an advantage over NL reasoners have a design such that participants produce a single intuition, and responding in line with this intuition biases decisions (Polonioli, 2018). Thus, the Stunted Intuitions Model easily explains the results of past investigations of the FLE. If intuitions are troublesome in a task, stunting them will improve performance on that task. If, however intuitions would be helpful in finding the correct response, stunting them would negatively affect the performance (see Polonioli, 2018 for similar claims). Further support for the Stunted Intuitions Model is provided by findings in moral decision-making reporting that using one's foreign language reduces all involved moral intuitions (Białek et al., 2019; Hayakawa, Tannenbaum, Costa, Corey, & Keysar, 2017; Muda et al., 2018). The Conflict Threshold Model alone cannot explain these previous findings, as an increased threshold for conflict detection would result in no difference as a function of language in a task that requires no conflict detection. We therefore think the Stunted Intuitions Model is a more promising candidate for explaining past findings on effects of thinking in a foreign language.

The present study cannot decisively distinguish between the Conflict Threshold and Stunted Intuitions Models. Both models explain our results with equal success because in syllogistic reasoning, conflict is binary; either it is present or it is absent. If, however, we studied the effects of foreign language reasoning in a task wherein the strength of the conflict was manipulatable, the Conflict Threshold Model would predict that the FLE disappears when conflict is at its highest because conflict will exceed even FL reasoners' increased threshold for conflict detection. One such family of tasks is base-rate problems (Bar-Hillel, 1980; Białek, 2017; Kahneman & Tversky, 1973; Turpin et al., 2020). In a base-rate task, participants must assess the probability of a person belonging to a particular group (i.e., a lawyer) while ignoring the salient, intuitive, stereotype information they have about the person (i.e., fitting to the stereotypical description of a lawyer). Critically, this

probability can be extremely low (i.e., there are 5 lawyers in our sample of 1000). This is an example of a high conflict problem, wherein the intuition clashes strongly with the base-rate information. Less extreme base-rates produce less conflict (i.e., 300 lawyers in our sample of 1000). The Stunted Intuitions Model would predict a FLE across high-conflict and low-conflict base-rate problems, as intuitions based on the base-rate will be stunted in either problem type. The Conflict Threshold Model, however, would predict that when the conflict becomes salient enough, even FL reasoners will be capable of detecting it and there will no longer be a difference between FL and NL reasoners. Some insight can be gained from the findings of a difference in moral judgments between FL and NL decision-makers in low-conflict but not in high-conflict moral problems (Chan, Xuan, Ng, & Tse, 2016; Costa, Foucart, Hayakawa, et al., 2014; Geipel et al., 2015). Preliminarily, this supports the Conflict Threshold Model, as the FLE only exists when conflict is low enough so as to be detected only by NL decision-makers.

7. Summary

Across three experiments, we show that bilinguals are worse at syllogistic reasoning in their foreign language. However, they are also equally biased by the irrelevant, belief-based dimension of syllogisms. This is in contrast to past findings regarding decision-making in a foreign language, which suggests that FL reasoners are protected from several common decision-making biases, especially if these biases are emotionally charged (Costa, Foucart, Hayakawa, et al., 2014; Gao et al., 2015; Hadjichristidis et al., 2019; Keysar et al., 2012). We provide evidence that this detriment in syllogistic reasoning is due to decreased ability to detect conflict between competing intuitions about conclusions' validity and believability. We propose two models which could account for this outcome: A Conflict Threshold Model, in which FL reasoners are unable to detect conflict which NL reasoners are able, but could detect conflict if it becomes so large as to pass their threshold; and a Stunted Intuitions Model, in which FL reasoners produce weakened intuitions, and this weakening eliminates conflict either by weakening any logical intuitions they have to a greater extent than their belief-based intuitions, or so much as to reduce their logical intuitions to zero. Further research is required to decide between these two models, or to develop a new, better fitting one.

CRediT authorship contribution statement

Michał Białek: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **Rafał Muda:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Funding acquisition, Resources, Project administration. **Kaiden Stewart:** Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Paweł Niszczota:** Formal analysis, Data curation, Writing - review & editing. **Damian Pieńkosz:** Investigation, Writing - review & editing.

Acknowledgements

We would like to thank Dries Trippas for providing us with the materials used herein and Evan F. Risko for helpful comments on previous versions of this manuscript.

The current project was financed by the resources of Polish National Science Centre (NCN) assigned by the decision no. 2017/26/D/HS6/01159 to MB. Work done by RM and DP was supported by the National Science Centre, Poland (NCN) under Grant PRELUDIUM 2018/29/N/HS6/02058. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104420.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109.

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning, 26*(1), 1–30.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233.

Białek, M. (2017). Not that neglected! Base rates influence related and unrelated judgments. *Acta Psychologica, 177*, 10–16.

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision making, 12*(2), 148–167.

Białek, M., Domurat, A., Paruzel-Czachura, M., & Muda, R. (2020). *Decision making does not benefit from using foreign language. Evidence from cognitive reflection effects on intertemporal choice. PsyArXiv.* https://doi.org/10.31219/osf.io/vhp86.

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.

Cavar, 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.

Chan, Y.-L., Xuan, G., Ng, J. C., & 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.

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

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS One, 9*(4), Article e94842.

Costa, A., Vives, M., & Corey, J. D. (2017). On language processing shaping decision making. *Current Directions in Psychological Science, 26*(2), 146–151.

Crane, N. (2016). *Debiasing reasoning: A signal detection analysis.* Lancaster University, UK: Unpublished Doctoral dissertation.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*(5), 428–433.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning, 20*(2), 169–187.

De Neys, W., & Białek, M. (2017). Dual processes and conflict during moral and logical reasoning: A case for utilitarian intuitions? In B. Trémolière, & J. F. Bonnefon (Eds.). *Moral inferences* (pp. 123–136). Hove, UK: Psychology Press.

De Neys, W., Moyens, E., & Ansteenwegen, D. V. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 208–216.

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science, 28*(5), 503–509.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*(3), 831–863.

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

Evans, J. St. B. T, Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*(3), 295–306.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Francis, G. (2014). Too much success for recent groundbreaking epigenetic experiments. *Genetics, 198*(2), 449–451.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking and Reasoning, 15*(2), 105–128.

Gao, S., Zika, O., Rogers, R. D., & Thierry, G. (2015). Second language feedback abolishes the "hot hand" effect during even-probability gambling. *The Journal of Neuroscience, 35*(15), 5983–5989.

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

Giner-Sorolla, R. (2018, January 24). Powering your interaction. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/.

Hadjichristidis, C., Geipel, J., & Surian, L. (2019). Breaking magic: Foreign language suppresses superstition. *The Quarterly Journal of Experimental Psychology, 72*(1), 18–28.

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 28–43.

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. *Psychology of learning and motivation.*

Vol. 62. Psychology of learning and motivation (pp. 33–58). Academic Press.
Hayakawa, S., Costa, A., Foucart, A., & Keysar, B. (2016). Using a foreign language changes our choices. *Trends in Cognitive Sciences, 20*(11), 791–793.
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.
Hayakawa, S., Lau, B. K. Y., Holtzmann, S., Costa, A., & Keysar, B. (2019). On the reliability of the foreign language effect on risk-taking. *Quarterly Journal of Experimental Psychology, 72*(1), 29–40.
JASP Team (2020). JASP (version 0.13.1) [Computer software]. https://jasp-stats.org/.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.
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.
Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*(4), 852.
Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning Memory and Cognition, 39*(4), 1265–1273.
Koen, J. D., Barrett, F. S., Harlow, I. M., & Yonelinas, A. P. (2017). The ROC toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behavior Research Methods, 49*(4), 1399–1406.
Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLoS One, 14*(3), Article e0211428.
Miozzo, M., Navarrete, E., Ongis, M., Mello, E., Girotto, V., & Peressotti, F. (2020). Foreign language effect in decision-making: How foreign is it? *Cognition, 199*, 104245.
Morsanyi, K., & Handley, S. J. (2012). Logic feels so good-I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning Memory and Cognition, 38*(3), 596–616.
Muda, R., Niszczota, P., Białek, M., & Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(2), 321–326.
Muda, R., Pieńkosz, D., Francis, K., & Białek, M. (2020). Author accepted manuscript: The moral foreign language effect is stable across presentation modalities. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820935072.
Muda, R., Walker, A. C., Pieńkosz, D., Fugelsang, J. A., & Białek, M. (2020). Foreign language does not affect gambling-related judgments. *Journal of Gambling Studies,* 1–20. https://doi.org/10.1007/s10899-020-09933-6.
Pavlenko, A. (2008). Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition, 11*(2), 147–164.
Pavlenko, A. (2012). Affective processing in bilingual speakers: Disembodied cognition? *International Journal of Psychology, 47*(6), 405–428.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34–72.
Polonioli, A. (2018). A blind spot in research on foreign language effects in judgment and decision-making. *Frontiers in Psychology, 9*, 227.
Raoelison, M. T. S., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition, 204*, 104381.
Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551–566.
Schmitz, C. (2010). *LimeSurvey [computer software]*.
Simonsohn, U. (2014, March 12). *No-way interactions.* http://datacolada.org/17https://doi.org/10.15200/winn.142559.90552.
Solcz, S. (2011). *Not all syllogisms are created equal: Varying premise believability reveals differences between conditional and categorical syllogisms (Unpublished doctoral dissertation)*. Ontario, Canada: University of Waterloo.
Šrol, J., & De Neys, W. (2020). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning,* 1–31.
Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought.* Yale University Press.
Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning, 24*(4), 423–444.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking and Reasoning, 20*(2), 215–244.
Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General, 147*(7), 945–961.
Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140.
Trippas, D., & Handley, S. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.). *Dual process theory 2.0* (pp. 28–46). Oxon, UK: Routledge.
Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: Complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning Memory and Cognition, 39*(5), 1393–1402.
Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: New indices for old effects. *Frontiers in Psychology, 5*, 631.
Trippas, D., Kellen, D., Singmann, H., Pennycook, G., Koehler, D. J., Fugelsang, J. A., & Dubé, C. (2018). Characterizing belief bias in syllogistic reasoning: A hierarchical Bayesian meta-analysis of ROC data. *Psychonomic Bulletin & Review, 25*(6), 2141–2174.
Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition, 45*(4), 539–552.
Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition, 133*(3), 586–600.
Turpin, M. H., Meyers, E. A., Walker, A. C., Białek, M., Stolz, J. A., & Fugelsang, J. A. (2020). The environmental malleability of base-rate neglect. *Psychonomic Bulletin & Review, 27*, 385–391.
Vives, M., Aparici, M., & Costa, A. (2018). The limits of the foreign language effect on decision-making: The case of the outcome bias and the representativeness heuristic. *PLoS One, 13*(9), Article e0203528.
Wickham, H. (2011). Ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics, 3*(2), 180–185.
Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*(4), 451–460.